Award Number:
W81XWH-13-1-0061

TITLE:
Novel Visualization of Large Health Related Data Sets - NPHRD

PRINCIPAL INVESTIGATOR:
William Ed Hammond, PhD

CONTRACTING ORGANIZATION:
Duke University

REPORT DATE:
November 2015

TYPE OF REPORT:
Final Technical Report

PREPARED FOR:
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:
Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
November 2015

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
25 Feb 2013 - 24 aug 2015

4. TITLE AND SUBTITLE

Novel Visualization of Large Health Related Data Sets - NPHRD

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-13-1-0061

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Hammond, William E; West, Vivian; Borland, David; Akushevich, Igor; Heinz, Eugenia, McPeck

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University
2200 W. Main St, Ste 710
Durham, NC 27705-4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

USA Med Research and Material Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT Using retrospective data queries to understand what information clinicians seek from health care data, we have identified data elements and are looking at combinations of data elements used in queries. We are developing various visualization techniques that can be used to present the informational content in large databases, expecting that visualization of this data will present or "discover" information without specific hypotheses. Groups of related data elements are incorporated into visualizations that allow a quick comparison of data from a large population, with the ability to view trends over time within a chosen category. We are exploring the ability to compress petabytes of health care data representing many data elements into various groups of related data presented visually with an interface that allows the user to interactively explore the data elements to understand big data from the perspective of an entire population, different disease groups, ages, and other variables. There is the potential to detect causal relationships between various sets of data, which when applied to military EHR data may lead to improved health care and resiliency in military personnel, assist the DoD in strategic decisions related to personnel, and save millions of dollars in health care costs.

15. SUBJECT TERMS
Visualization, health care data, big data

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	238	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

1. INTRODUCTION

With the growth of Electronic Health Record (EHR) data and other related healthcare databases, there is a growing interest in using this data to understand what information and knowledge the data represent. Visualization offers an opportunity to explore and understand large data in unique and novel ways, permitting one to view data without the bias of an a priori decision of what is important. We hypothesize that data visualization is more effective than traditional methods of data exploration, and that the type of visualization is highly dependent on the types of data and nature of the queries and what someone is trying to learn from the data. The aims of this research were to use retrospective data queries to understand what information clinicians seek from health care data; identify what data elements and mixtures of data classes (laboratory data, demographic data, problems, therapies, physical examination data, or imaging data) are used in queries and what methods are used to analyze query results; and create a matrix of data visualization methods used with specific data elements from multiple classes and test visualization of mixed data classes. We used various visualization techniques to present results of data queries and developed and evaluated the informational content from a large amount of data to see what presented or "discovered" itself without a specific hypothesis.

2. KEYWORDS

Visualization, health care data, big data, data elements, temporal analysis, visual analytics, multivariate visualization, radial coordinates, path map, queries

3. OVERALL PROJECT SUMMARY

Our research start date was 25 February 2013. In the months previous to this, we were proactive due to the short duration of the funding (18 months) and had just received approval from Duke's IRB to commence the research. Although we submitted our documentation to the Human Research Protection Office (HRPO) as one of our first tasks, we did not receive HRPO approval to begin working with clinical data until the end of September 2013. As a result, we were delayed in completing our project milestones and associated tasks. A no-cost extension with slightly revised milestones, tasks, and timeline was submitted and approved, which extended our research to the end on 24 August 24 2015. For the purposes of this Final Report, we are reporting on the revised set of Milestones and Tasks. Section 3 summarizes the results of our research per the nine Project Milestones and their associated tasks

Milestone 1. Obtain access to queries of Duke's Clinical Data Warehouse

At Duke we use an on-line query system called DEDUCE (Duke Enterprise Data Unified Content Explorer)[1] to access data in the Decision Support Repository, which consists of hundreds of tables, some with hundreds of millions of rows of data, collected from over 3.3 million patients at Duke. DEDUCE was operationalized in 2008 and upgraded numerous times during the next three years. By 2011, it was recognized within the Duke medical community for its value in abstracting information from the Data Repository. Researchers can query over 10,000 data elements and refine the query to facilitate exploration of aggregate clinical data in support of operations, quality, and research. Output from the queries is in the form of common-separated values (CSV) files, ASCII files, Excel files, or simple graphs. Every query is saved on a Duke server. We had the unique opportunity to use the clinical data from these retrospective data queries of Duke's Clinical Data Warehouse for our research.

Our first milestone was to obtain access to the archived queries, conduct an in-depth review to identify what information clinicians seek from the data, and determine what data elements are used. While we waited for HRPO approval to begin our review, we obtained a count of terms used to filter subject areas using DEDUCE. With data from May 2010 through April 2013, we explored ways to visualize this metadata. We developed a prototype interactive force-directed node and link network visualization using the Processing programming environment to visualize the DEDUCE queries (Figure 1).

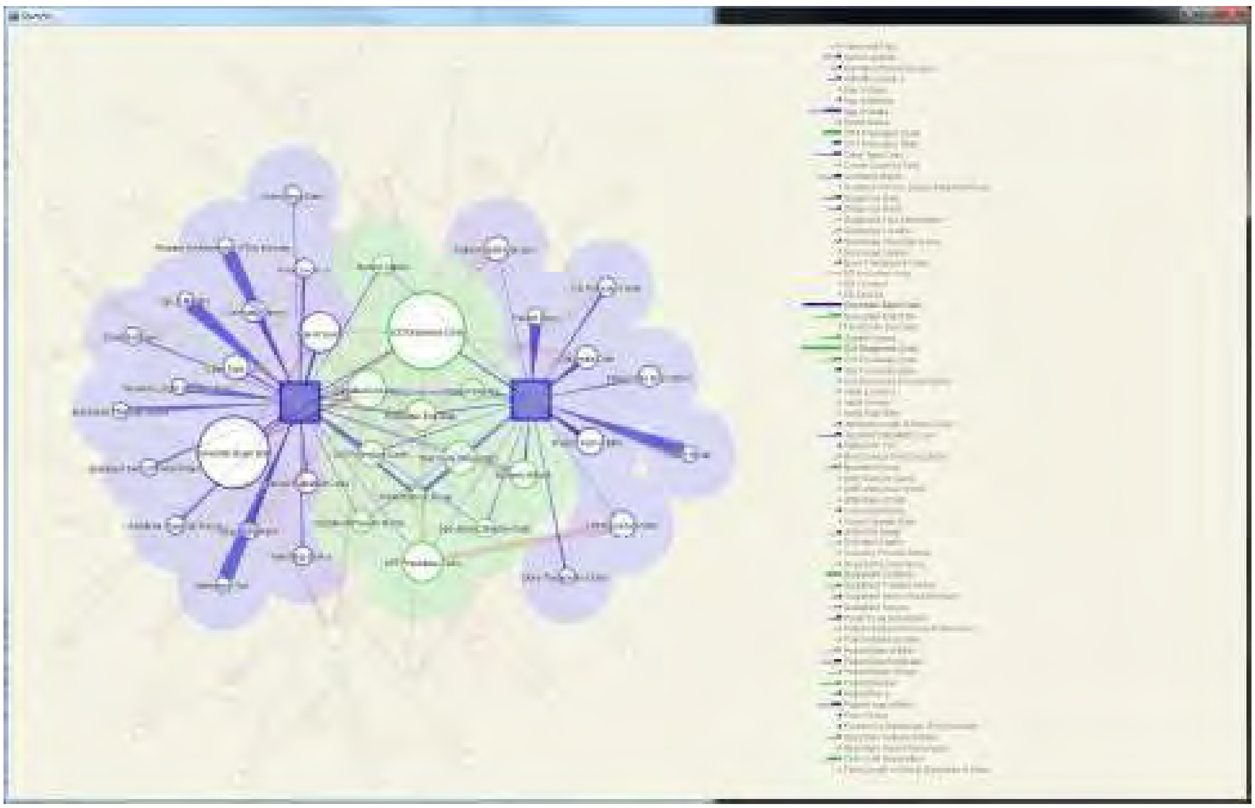


Figure 1. Force-directed layout visualization of DEDUCE queries.

This visualization shows data elements as circles, with the size of each circle representing how often it was queried. Squares represent de-identified system users (only the top-two users are shown in Figure 1), and the size of each square represents the number of queries made by that user. Links between nodes represent how often each element was used together in a series of queries, with each end scaled according to the relative importance at each end of the link. Links between circles and squares represent how often each user made a query on each data element. Nodes are placed via a force-directed layout based on the overall strength of each link.

As shown in Figure 1, when the two users are highlighted this in turn highlights nodes connected to those users while de-emphasizing all other nodes. Nodes that are connected to both users are highlighted in green, whereas nodes that are connected to just one user are highlighted in blue. A full list of data elements is shown to the right, with horizontal lines representing the number of times each data element was used across all queries (equivalent to circle size), and the number of other data elements connected. The user can interactively select nodes via the node-link diagram or the list of data elements. A video showing selection of the Admitting Physician Name and Discharge Physician Name nodes can be found at http://wwwserver-1.renci.org/~borland/movies/QueryVis_01_c3.avi

From this visualization we can see that the data element most queried by all users was ICD-9 diagnosis code, represented by the largest circle in the green in Figure 1. One of the users queried ages in months only, while the other queried age in years only, leading us to conclude that one user was a pediatrician and the other oriented to adults. Although this was not something we had planned to do in our research, it provided us with a great deal of information about the types of data elements queried by DEDUCE users and frequency of their use. Merely looking at the tables of the data used, we would not have been able to identify these visual results, supporting our hypothesis that looking at data visually is more effective than traditional methods of data exploration. It also helped us address Milestone 3, which was to identify data elements used in queries.

When we received HRPO approval at the end of September, we then requested retrospective query data from 01 Jan 2011 through 31 Jul 2013. We used only half of these queries, however; the volume of data from the queries surprised us. It represented a significant amount of information as we began working with this big data. Copies of the queries were transferred to a secure workspace on a protected Duke server for our research team to analyze.

Milestone 2. Develop classification for queries

(a) Identify early use of data queries
(b) Based on the reason for the queries, group them accordingly.
(c) Obtain access to AHLTA de-identified data, and using work from the Duke queries and classes, compare for similarities and differences and revise classes as needed.

Because of the delay in obtaining data from the DEDUCE queries, which was to include interviews with users to determine how they were using query data (a task in Milestone 4), we in the interim developed a survey of those people within Duke who were identified as DEDUCE users. The survey was built using REDCap (Research Electronic Data Capture), a web-based application for building and managing electronic data capture. At the end of October 2013, an email request to participate in the survey was sent to 482 users who, within the two years previous years, had completed the training course required to use DEDUCE. A reminder email was sent two weeks later; the survey was closed at the beginning of December 2013. Participants were anonymous to us unless they volunteered to participate in a later evaluation, with a question asking for their name and contact information if volunteering (See Appendix O). A total of 61 people completed the survey, a 12.7% response rate. Of the 61 responders to the survey, 34 people provided us with contact information and indicated a willingness to be contacted individually as we evaluated different visualizations (this is described in Milestone 4).

Respondents were asked to identify the reasons they ran queries. The most frequent aims for conducting queries were related to grant preparation, determining the prevalence of a particular population or getting other information for a grant (47.4%), quality improvement (46%), outcomes (33%), and to see if there were enough patients who would meet inclusion/exclusion criteria for participation in industry-sponsored clinical trials (28%). Of the additional clinical and non-clinical reasons for queries, 28% of the responses were also related to research activities. When asked to approximate the number of times the respondents initiated queries in the previous two years, the majority selected 1-4 queries (39%), with 29% selecting 15-19 queries (29.8%); 12% conducted queries more than 20 times. Respondents were asked to identify the types of information typically sought in a query, selecting all that applied. Respondents received the output from their queries as Excel tables (77%), ASCII files (18%), or CSV files (5%). Of those participating in the survey, 77% were satisfied always or most of the time with the information obtained as a result of the queries. Results are noted in Table 1.

Query Data Requested	Percent (%)
Diagnoses	83
Demographics	75
Encounters	47
Procedures	46
Medications	39
Laboratory data	35
Physicians	28
Imaging data	23
Vital signs	16
Text for analysis	12
Geospatial data	9
Device information	5

Table 1. Percentage of users and type of data most frequently requesting using DEDUCE queries.

Throughout the first 18 months of our research we examined the queries, the structure of the queries, and the type of data requested, in addition to the apparent purpose for the queries. We knew from our force-directed network using DEDUCE query terms described in Milestone 1 that ICD-9 codes and the date of encounters were the most frequently queried data elements, which was consistent with our survey data above. Using visualization techniques in R, we looked at the relationships between the data elements used in queries, how often each data element was used in a series of queries, and how often each user queried each. This process helped determine the most common relationships between data elements used in queries. (Note: Refer to article in Appendix A for additional details.) We realized that the DEDUCE queries were used almost exclusively for two reasons: the most frequent use was for research (with numerous variations of the kind of research and purpose, e.g. cohort identification or counts of patients to determine feasibility of participation in a clinical trial); the second was for quality improvement. The only exception was to identify patients with specific data to be used for medical education. Therefore, categorization of the queries would be (1) research, (2) quality improvement, and (3) other.

Our final task to complete Milestone 2 was to obtain access to AHLTA data that we had planned to use for comparing similarities and differences with Duke's retrospective query data. We also had planned to use AHLTA data to look at people with a diagnosis of PTSD. Instead we were given access to TATRC simulated data being modeled to mimic actual data. We found that an early set of 10,000 patients appeared to be primarily from traumatic brain injury patients, which restricted what we could do with the data: we wanted to view data without the bias of an a priori decision of what is important. A second data set of 1 million patients was provided that we hoped would be more random.

Using the multiple tables from the synthetic dataset, our statistician incorporated the various combinations of data elements into groups so they could be used in more efficient ways. The frequency distributions for all variables in the dataset were structured for further analyses and visualization. We created four specific tables with patient-based, encounter-based, medication-based, and lab-based information to avoid using a measurement for a variable several times. For the patient-based data we created a table with all disease onsets with information about the date at onset. Both datasets (10,000 patients and 1 million patients) were processed.

Working with the synthetic data was quite restrictive and problematic for us to use: the data were more a set of numbers with labels than random data. For example, all laboratory values for HgA1c were identical. In addition, everyone was listed with their home address around the Bethesda area, data we believe would to be of significance when visualizing large amounts of data: providers in a specific location may differ in the tools they use for diagnosis and in treatment plans, which we would not be able to see using the synthetic data. PTSD occurred in 8,565 patients in the million patient data, and had comorbidity with 75 out of 105 problems in the data. However, there appeared to be two completely disjointed sets of patient problems in the data. In Figure 2, the cluster of problems in the center-right of the images, mostly pink, including PTSD in red, and is completely disjointed from the more spread out cluster of blue nodes in the lower-left, meaning no patient with a problem in one cluster also had a problem in the other. PTSD is also not connected to the other blue nodes in the middle of the center-right cluster. The separate cluster of blue nodes corresponds exclusively to traumatic events such as amputations, disarticulations, and concussions. In Figure 3, the image shows the odds ratios for co-morbidity for PTSD with all other problems in the data base. The problems at the bottom, with a value of zero, correspond to these blue nodes in the image above. In addition to traumatic events, PTSD is also not associated with schizophrenia, paranoia, depression, bipolar, or anxiety.

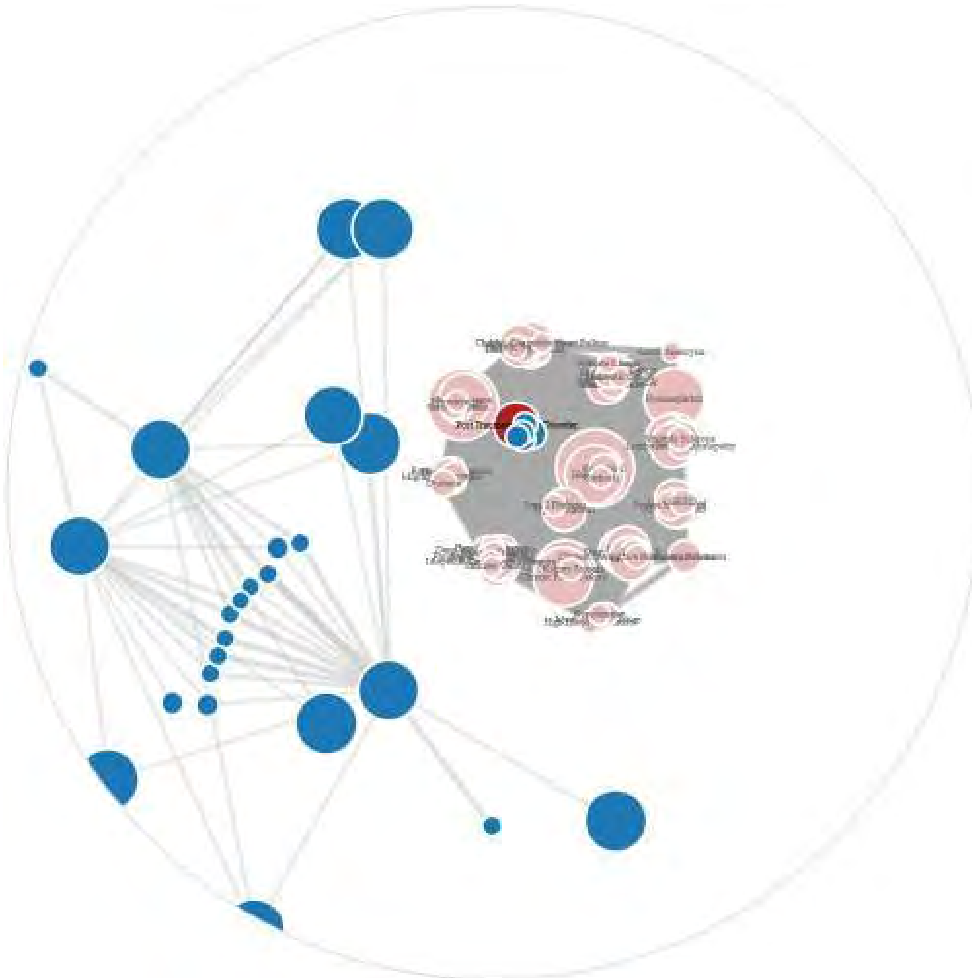


Figure 2. PTSD data for synthetic data. The cluster of problems in the center-right of the images, mostly pink, including PTSD in red, and is completely disjointed from the more spread out cluster of blue nodes in the lower-left. This shows that no patient with a problem in one cluster also has a problem in the other. PTSD is also not connected to the other blue nodes in the middle of the center-right cluster. The separate cluster of blue nodes correspond exclusively to traumatic events such as amputations, disarticulations, and concussions.

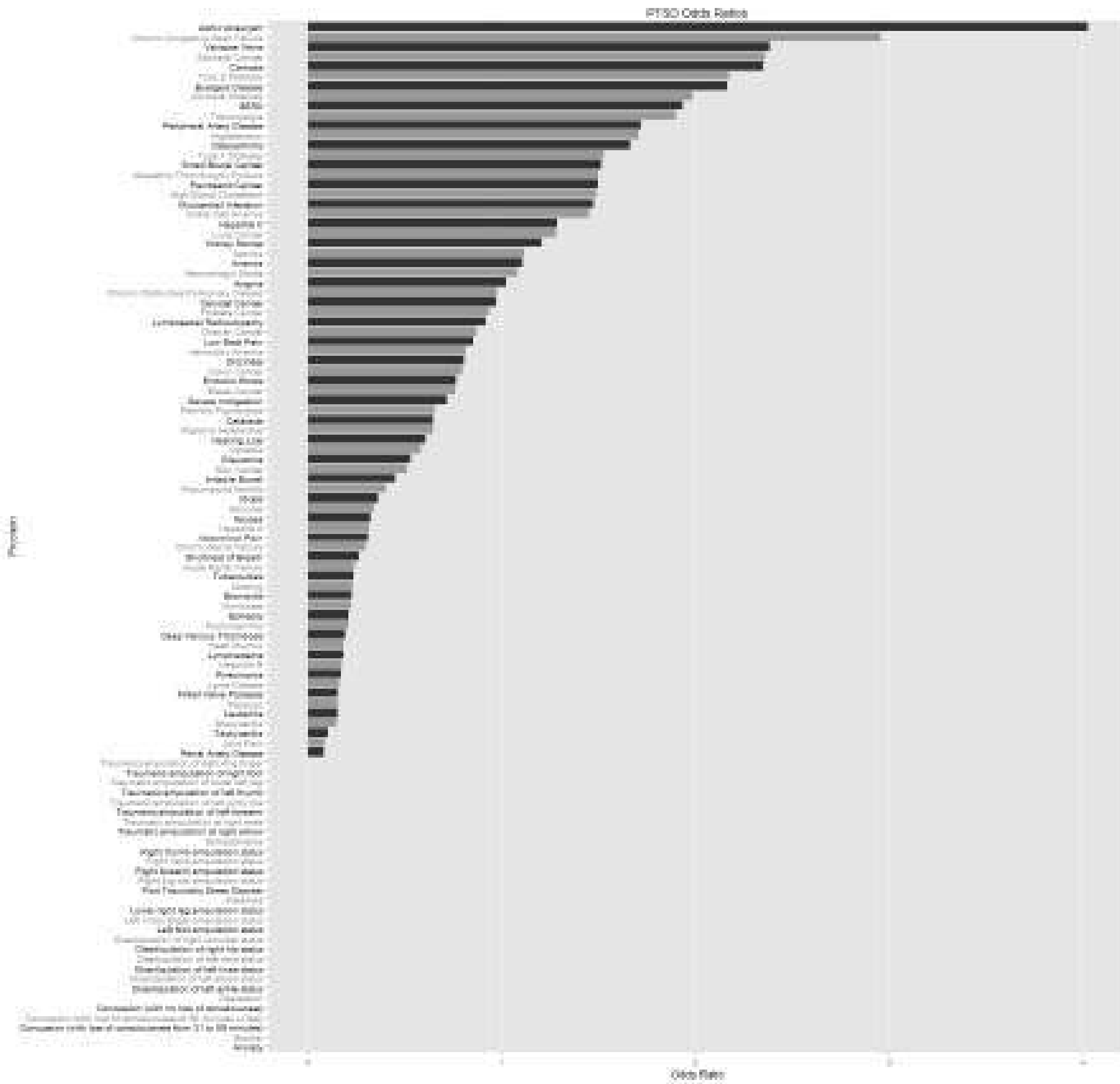


Figure 3. Odds ratios for co-morbidity od PTSD with all other problems in the data base. The problems at the bottom, with a value of zero, correspond to the blue nodes in the image above. In addition to traumatic events, PTSD is also not associated with schizophrenia, paranoia, depression, bipolar, or anxiety.

We talked to the developers of the synthetic data and found that data are constructed based on input from medical experts who provide the developers with likely symptoms, diagnoses, and therapies. The developers could customize the system to generate data for us, but it is a rules-based generation of data that would require specific input by us as to what types of data we wanted (e.g., 30% of all patients with PTDS, with 50% of these having comorbidities of other mental illnesses). Because our research is based on the premise that visualization a priori of large data will lead to discoveries not previously known, the synthetic data did not lend itself to our research. Our fallback was to use data from the Duke data warehouse and run queries that would allow us to test our ability to visualize data and discover new knowledge as described in Milestone 9. We therefore were not able to compare AHLTA data with Duke data for similarities and differences that might revise our grouping of queries.

Milestone 3. Identify data elements used in queries

(a) Identify which queries should be most meaningful to include in our analysis and the individual researchers associated with those queries.
(b) Group data elements into classes (e.g. laboratory data, demographics, medications).

The amount of data in the DEDUCE queries was immense, which made a manual review overwhelming beyond randomly looking at the data elements usually requested to see if a classification schema could assist in visualizations. DEDUCE, our query tool, constrains the types of data that can be queried, and we found that more seasoned users of the system query more data than newer users, which may be a result of becoming more familiar with the nuances of the tool itself. The force-directed visualization of data elements we did at the beginning of our research is, interestingly, perhaps the most useful way to understand the data elements used in the query. Using the force-directed network visualization allowed us to interactively explore the departments doing queries, the data elements that were queried, and see what is used most frequently and the correlation with other data elements. We also validated the importance of interactive visualization. The complexity of the presentation of data requires the use of additive measures of points of interest, which is difficult to understand without the ability to interactively highlight certain types of data to see various correlations and highlight the relationships among the data elements.

A classification schema was more difficult to develop than we had expected, and we considered many combinations. For example, while medications might be one category, a number of medications for adults and children differ, raising the issue of perhaps using a less inclusive classification, e.g. adult medication and pediatric medication. Within medications, we considered whether it is better to use drug classes to be more specific. After an exhaustive review of the retrospective DEDUCE data queries and their structure, we concluded data elements in data queries at Duke are already grouped into classes by the data dictionary used in the tool. The broad categories are noted in Table 2. These can be future categorized into subcategories, and some of these are further subcategorized (e.g., Patients is a category, Social History is a subcategory, and subcategories of Social History are Alcohol, Tobacco, Illicit Drug Use, and Birth Control).

Category	Sub-category	Category	Sub-category
Patients	Demographics	Results	Discrete Lab Results
	Social History		Culture Results
	Allergies		Vitals
Patient Geography	Address information	Orders	Pharmacy
	Demographics (census)		CPOE Orders
Encounters	Hospital	Radiation Oncology	Diagnosis and Treatment
	Clinic		Radiation Therapy
Diagnoses		Medications	Patient Medications
Procedures	ICD Procedures		
	CPT Procedures		

Table 2. Categories and first sub-category of Duke queries using DEDUCE.

We also found that as we explored the query data and found interesting information through its visualization, we wanted to look at various relationships of data elements that were important to understanding what we were seeing. The research itself then guided us in determining the classes we needed. For example, using all ICD-9 codes was ineffective due to the number of codes (and this will only increase with ICD-10 codes). Therefore, using the primary classes developed within the International Classification of Diseases (ICD) schema was broad enough for us to determine the relationship of ICD-9 diagnoses for people diagnosed with PTSD, and we condensed the longer codes into the set of primary classes (see Table 3). It would be very interesting to have access to another institution's data warehouse to know if and how data elements are grouped into classes and how a query of this data differs from the Duke query tool.

	015 ICD-9-CM Diagnosis Codes
001-139	Infectious And Parasitic Diseases
140-239	Neoplasms
240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
280-289	Diseases Of The Blood And Blood-Forming Organs
290-319	Mental Disorders
320-389	Diseases Of The Nervous System And Sense Organs
390-459	Diseases Of The Circulatory System
460-519	Diseases Of The Respiratory System
520-579	Diseases Of The Digestive System
580-629	Diseases Of The Genitourinary System
630-679	Complications Of Pregnancy, Childbirth, And The Puerperium
680-709	Diseases Of The Skin And Subcutaneous Tissue
710-739	Diseases Of The Musculoskeletal System And Connective Tissue
740-759	Congenital Anomalies
760-779	Certain Conditions Originating In The Perinatal Period
780-799	Symptoms, Signs, And Ill-Defined Conditions
800-999	Injury And Poisoning
V01-V91	Factors Influencing Health Status and Contact With Health Services
E000-E999	Supplementary Classification Of External Causes Of Injury And Poisoning

Table 3. ICD-9 Classification System. From http://www.icd9data.com/2015/Volume1/001-139/default.htm.

When we began our research, we had expected to visualize data from data elements at a very granular level. As time evolved and we worked with this big data, however, we found that it is not realistic to drill down to specifics at this point in time. There are numerous challenges to be addressed before this can be considered. Two were apparent to us immediately. First, the size of the retrospective query data files created a storage problem that required special approval to store our data on a Duke server that could handle the capacity. Second, we also had to consider the size of the data files using the visualization programs that we used, how quickly the computations could be completed, and how stable the programs were to handle this large data. There are efforts underway in working with big data (e.g., Hadoop) and we will consider some of these innovations in our future research.

<u>Milestone 4</u>. Explore alternative visualization methods of the data. Clinicians will use a Follow-up Questionnaire to compare alternate visualizations of the data to the original presentation of the query data.

(a) Explore visualization methods previously applied to health data.
(b) Conduct semi-structured interviews with clinicians to determine how the user intended to use the data from the query, the relevance of the query, and the clinicians' satisfaction and use of the information derived by the query.
(c) Develop visualizations of retrospective query data using standard visualization techniques and novel visualization techniques to share with clinicians
(d) Develop a Follow-up Questionnaire using a 5 point Likert scale to be used in researcher evaluation of different visualization methods.
(e) Combine interview data with Questionnaire results to evaluate clinical relevance of the visualization methods.

The first task for completing Milestone 4 was to explore visualization methods previously applied to health care. Following the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) Statement[2], a systematic electronic review of the literature was conducted between May and July 2013 to investigate the use of visualization techniques reported between 1996 and 2013. A review using MEDLINE and Web of Knowledge was supplemented with citation searching and a grey literature search. Reference lists from highly relevant articles were also reviewed to find additional articles. Broad key words and search terms were used to assure a comprehensive document search. A matrix was developed for reviewing and categorizing all abstracts and to assist with determining which should be excluded in the review. Articles were excluded if they related to genetics, animals, environment, population health, primarily related to the technical aspects of visualization or position papers, or did not describe specific techniques used for the visualization.

Eighteen articles were included in the qualitative review. Although there is increasing interest in visualization using health care data, in particular population data from EHRs, its use is limited. A manuscript on this review and findings was successfully submitted to JAMIA, first published online 23 October 2014 and in print format in March 2015 [3] (See Appendix H article that includes the schema adapted from the PRISMA group.). The article was presented to the JAMIA Journal Book Club 5 February 2015 (see Appendix I).

Our second task to complete this milestone was to conduct semi-structured interviews with clinicians to determine how the user intended to use the data from the query, the relevance of the query, and the clinicians' satisfaction and use of the information derived by the query. We described the survey that was conducted instead as we waited for approval to begin working with data (see Milestone 2).

Another task in Milestone 4 was to develop a variety of visualizations of retrospective query data using standard visualization techniques and novel visualization techniques to share with clinicians. A review and evaluation of the visualizations was completed by ten respondents familiar with queries of the Duke EHR. A Follow-up Questionnaire using a 5 point Likert scale (See Appendix P) was developed using REDCap, which was completed by participants to evaluate and rank each visual display as a scripted explanation of each was provided by the interviewer. We had a large portfolio of different visualization methods, which including interactive visualizations we completed using three retrospective DEDUCE data sets. The evaluation included brief hands-on use of the parallel sets and radial coordinates visualizations we developed (discussed in Milestone 7).

We divided the various visualizations into three distinct sets based on the type of data we were using in the visualization.

- **Set 1** consisted of 600 rows of data with the following data elements: unique encounter ID, admit date, DRG code or ICD code, DRG or procedure description, and inpatient length of stay in days. There were a total of 86 unique DRGs. We divided visualizations into categories based on three main topics: (a) inpatient length of stay, (b) trends over time, and (c) number of DRGs or number of times a particular event occurred over time.
- **Set 2** contained the vital signs for 240 patients with the following data elements: member ID, encounter ID, height, weight, systolic blood pressure, diastolic blood pressure, pulse, respirations, and temperature. The visualizations represent (a) the number of visits over time and (b) how vital signs change over time.
- **Set 3** consisted of 2,940 patients with the following data elements: patient ID, death status, ethnic group, gender, diabetes indicator, hypertension indicator, race, religion. This set was divided into four categories depending on the number of variables used (from one to ≥four) and four interactive visualizations.

We were over-ambitious in the number of visualizations we expected to cover during each hour-long session, and after the first evaluation eliminated half, assuring that at least one of each kind of visualization was included, as follows.

1. Bar graph
2. Radial Coordinates
3. Bipartite graph
4. Sankey diagram
5. Box & Whiskers
6. Scatterplot
7. Bubble graph
8. Scatterplot distribution
9. Heatmap
10. Slope/Best of fit
11. Lines graph
12. Stacked bar graph
13. Marimekko Chart
14. Stream graph
15. Parallel sets

The respondent had a significant effect on the way the interview was conducted and the time involved, with each individual's orientation to facts versus trends, populations versus individuals, concrete versus abstract reasoning, familiarity with the visualization used, and visual thinking all contributing to interpretation of the visualizations. A comment from an evaluator who is a radiologist and the only participant who assessed all of the visualizations demonstrates how experience plays a large part in the response to visual data, stating that she is used to looking for patterns in the visual images she sees. The importance of usability was evident as the evaluators commented about such things as legends, labels, color and size of graphical representations, placement of axes, ease of understanding at a glance, and amount of data represented.

We paced the evaluation to assure each evaluator had hands-on time with the interactive visualizations. Time to explain what the visualization represented and the nuances for each interactive visualization was limited, however, and the smaller size of the laptop screen prohibited the best visual display and ease in seeing the data. Overall, however, the ability to act interactivity with data was well-received, and respondents commented that the visualization provided more data, specific data was easier to find, and information could be grasped better to show trends. A manuscript regarding our findings from these evaluations is in process. Further research on usability with a structured interview is one of our future research goals.

<u>Milestone 5</u>. Modify or revise classification and data elements of queries based on analysis of the relevance of the visualization methods.

After analysis of the visualization methods in Milestone 4, it was evident that classification of data elements for queries is affected by too many confounders to be able to develop a precise classification. People who participated in our evaluation were from a variety of healthcare backgrounds including researchers, clinicians, administrators, and teachers. The single variable they had in common was their interest in healthcare data. The evaluation of a visualization's usefulness and the respondents' ability to quickly find information in a specific visualization varied, however. We could find no single explanation to identify why there was such a variety in the evaluation of each visualization. It was not uncommon for a respondent to make positive or negative comments when looking at a particular visualization, yet ranked it higher or lower than the comments indicated. We conclude that static visualizations a health care professional uses best are those they are most familiar with or display data as trends for large amounts of data, or simple charts for small amounts of data. We further conclude that people working in health care are receptive to interactive visualizations, see the value in interacting with larger amounts of data than they are used to seeing, and that training time to effectively use the visualization will be important to the success of interactive visualizations.

<u>Milestone 6</u>. Create a matrix of best visualization techniques.

(a) Explore ways to mix different types of data in visualization.

We have struggled with this milestone as our own research team has had difficulty quickly understanding some of the data visualizations we have used. As these visualizations have been refined and our team had grown to understand what each represents, we have gradually gained a deeper appreciation for the nuances of visual representation of data, recognizing that in spite of the 30 months we have spent on this project, determining the best visualization is most likely an individual preference, not a conclusion that can be scientifically proven. We also did not begin our research fully appreciating the amount of knowledge that can be gained from studying what we were seeing as data were visualized, which in turn pushed us to ask additional questions of the data and question if the visualizations could be refined to answer those questions.

We have used several different visualization techniques and have found there are a number of challenges to all interactive data. These include:

(1) how to manage massive amounts of data;
(2) how to best display temporal data;
(3) what can be done to normalize complex data and differentiate between categorical and numerical data;
(4) how to manage clutter in a visualization;

(5) what to do about missing or inaccurate data; and
(6) how quickly someone can understand a visualization, how to use it, and its value.

The data itself is a significant variable in determining the "best" visualization techniques. The Duke data we used is from a unique Duke query system (DEDUCE) that allows researchers to pull data from the data warehouse. We know that there are many variations of data elements associated with every health care system's data warehouse and had hoped to use AHLTA data to evaluate whether our visualizations could be seamlessly applied to another set of data. Although we feel confident that our tool is generalizable, we cannot say this with certainty. We hope to be able to do this in future research, continuing to pursue how or if a matrix can be developed for use by others wanting to use the best visualization technique for a specific data set.

Although determining a complete matrix of "best" visualization techniques proved difficult, we have experimented with various visualization techniques, and ways to mix different types of data. Our radial coordinates technique, described in more detail in Milestone 8, uses data-type-dependent axis distribution visualization and curve spreading to mitigate some of the typical problems encountered when combining continuous and categorical data in parallel-coordinates techniques. It also incorporates linked views to provide different views of the data that can lead to better insights. Below we describe visualization techniques developed to handle temporal and textual data.

Temporal visualization of diabetes data

We have developed a tool for the visualization of disease trajectories over time based on the parallel sets[4] visualization technique (Figure 4). In this visualization we align patients diagnosed with diabetes by death at the right, and visualize the paths of the HbA1c values of groups of patients going back in time from death, categorizing their lab values as *Normal*, *Borderline*, *Controlled*, and *Uncontrolled*. Data for the 535 patients is sampled every 6 months, and the users can control the sampling rate of the visualization above that. The visualization shows how much variability there can be in HbA1c values, and also shows a trend towards normalization at death; the green *Normal* section of the vertical axis at death is much larger than at the previous sample two years before death. The user can select any part of any path through the data to highlight that data and show the number of patients and percentage of the total represented by that path. Figure 4b highlights all patients with *Normal* HbA1c values at death, illustrating how many patients moved towards *Normal* at death.

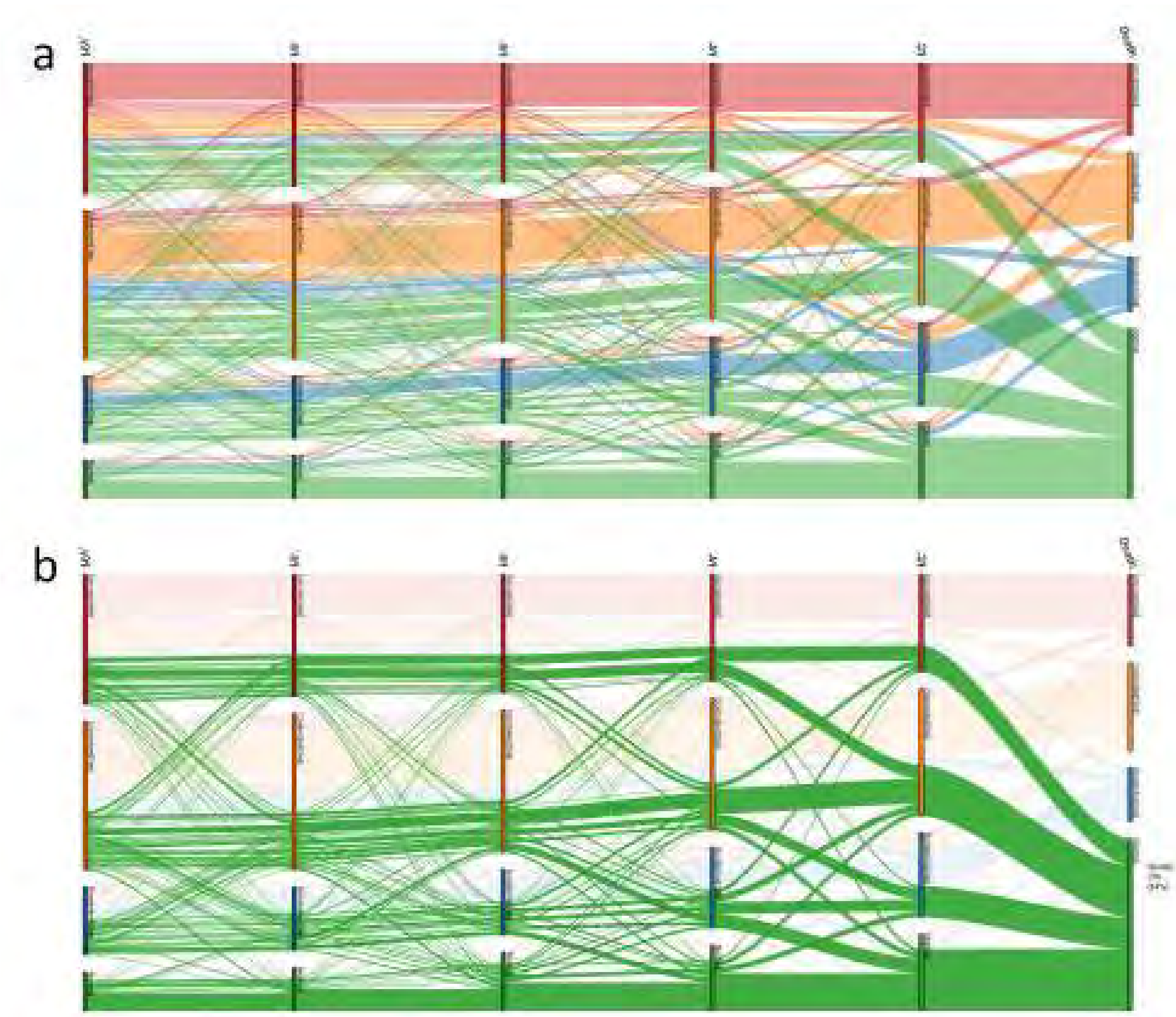


Figure 4: Parallel sets visualization of HbA1c values over time in diabetic patients.

We have added various controls to this visualization, such as the ability to show trajectories moving forward in time, such as Figure 5, where we compare the forward trajectories of patients who remained *Uncontrolled* (red) from -10 years to -4 years to those that remained *Normal* (green) from -10 years to -4 years.

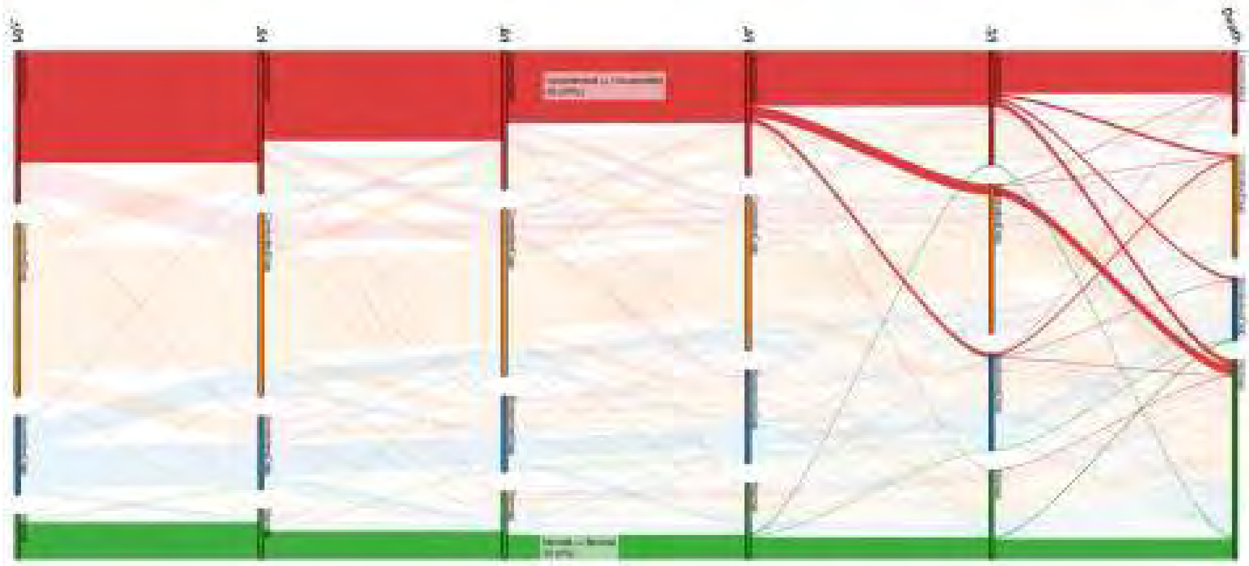


Figure 5: Comparing forward trajectories of *Uncontrolled* (red) and *Controlled* (green) starting 4 years before death.

We have also added the ability to incorporate non-temporal data, such as age, gender, and race, with our temporal visualization. Parallel sets is designed to enable the visualization of categorical data, so numeric data, such as age, must be transformed to an (ordered) categorical variable (e.g. decade). Categorical data can then be incorporated as additional vertical axes (Figure XXX). More details about this visualization, can be found in Appendix F.

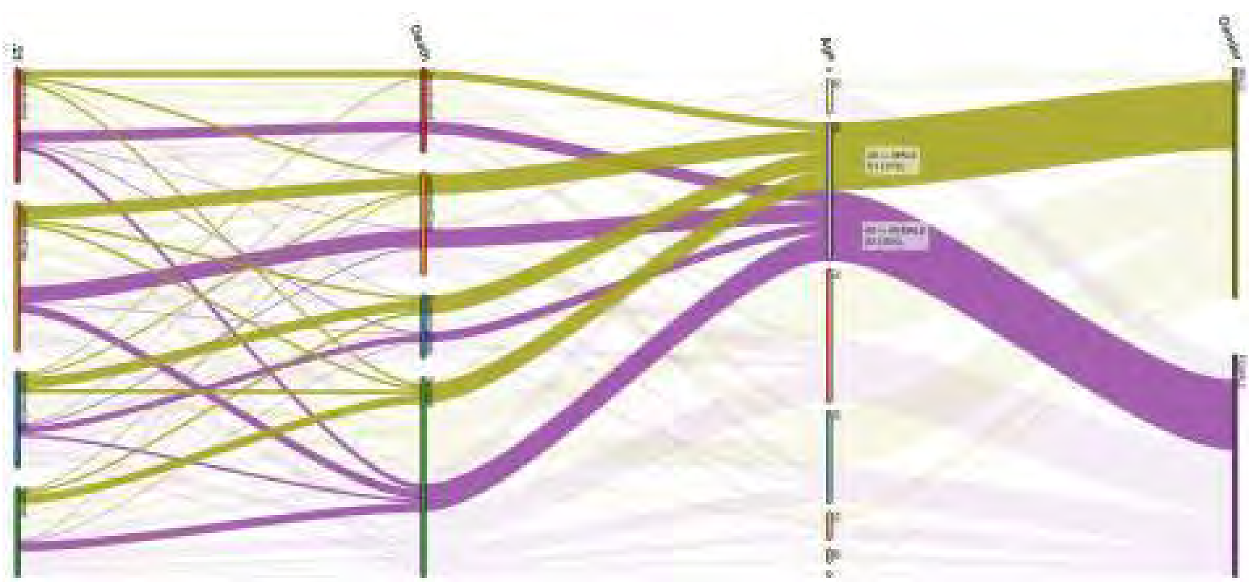


Figure 6: Adding numeric (*Age*) and categorical (*Gender*) axes. Here we compare the HbA1c trajectories between death and two years prior to death for males (yellow) and females (purple) in their 80s.

One of the advantages of the parallel sets temporal visualization is its ability to represent large numbers of patients, as paths are drawn for groups of patients with the same trajectory, with path width encoding the relative number of patients with that path. A disadvantage of this visualization is the increased splitting of paths as more temporal samples are added, making it difficult to follow individual paths (Figure 7).

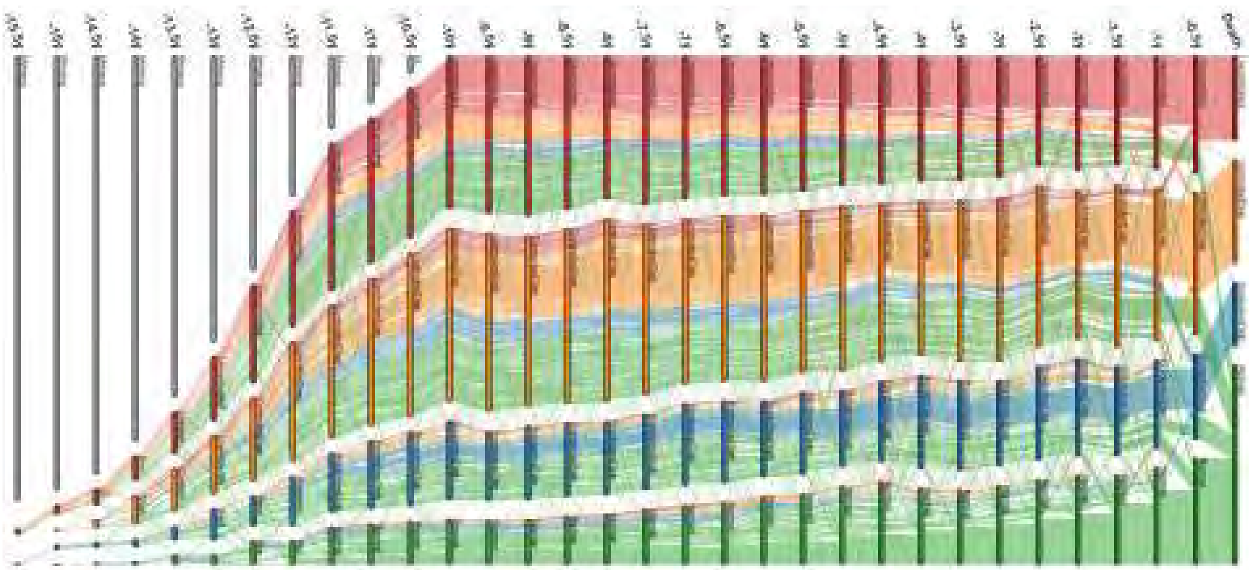


Figure 7: Parallel sets temporal visualization showing 15 years of data, sampled every 6 months.

We therefore developed another temporal visualization technique, path maps, to enable enhanced perception of individual paths while still showing overall trends in the data. Our path map visualization is based on the heat map technique[5] consisting of a 2D grid with each data cell assigned a color based on value, with modifications specific for temporal data.

Figure 8 provides shows our path map visualization applied to the same data used for the parallel sets temporal visualization. Figure 8a shows the main path map visualization. Each row is an individual patient, and each column is a temporal sample point (every 6 months in this case), with data aligned by patient date of death on the right. The ragged left edge shows that patients had differing temporal ranges of data collection. HbA1c value categories are mapped color, with *Normal* as pale yellow through *Uncontrolled* as dark red. Figure 8b shows the distribution of each category across all temporal samples per patient, Figure 8c shows the distribution of each category across all patients per temporal sample, showing overall trends in the data, and Figure 8d shows the distribution of each category across all patients and temporal samples. A key feature of the path map visualization is how patients are sorted vertically. The user can click in any column to select that temporal sample, causing all patients to be sorted by their value at that sample. In Figure 8a, the user has selected a sample 5 years before death, highlighted by the black vertical lines. Black horizontal lines help the user visually segment the sample before and after the selected sample, with the pattern of colors in each segment showing the distribution of values. Various sorting methods can be used after sorting by the selected temporal sample to bring out different features of the data. Figure 8a uses a weighted average around the selected sample, whereas Figure 8e sorts first by the selected sample, then moving backwards from the last sample. This method enables us to clearly see, for a given data value at a given time point, where those patients ended up. This data is emphasized by the thin colored lines within the highlighted column, which are colored by value at the last sample. For example, we can see that many more *Uncontrolled* patients ended up *Normal* than vice versa.

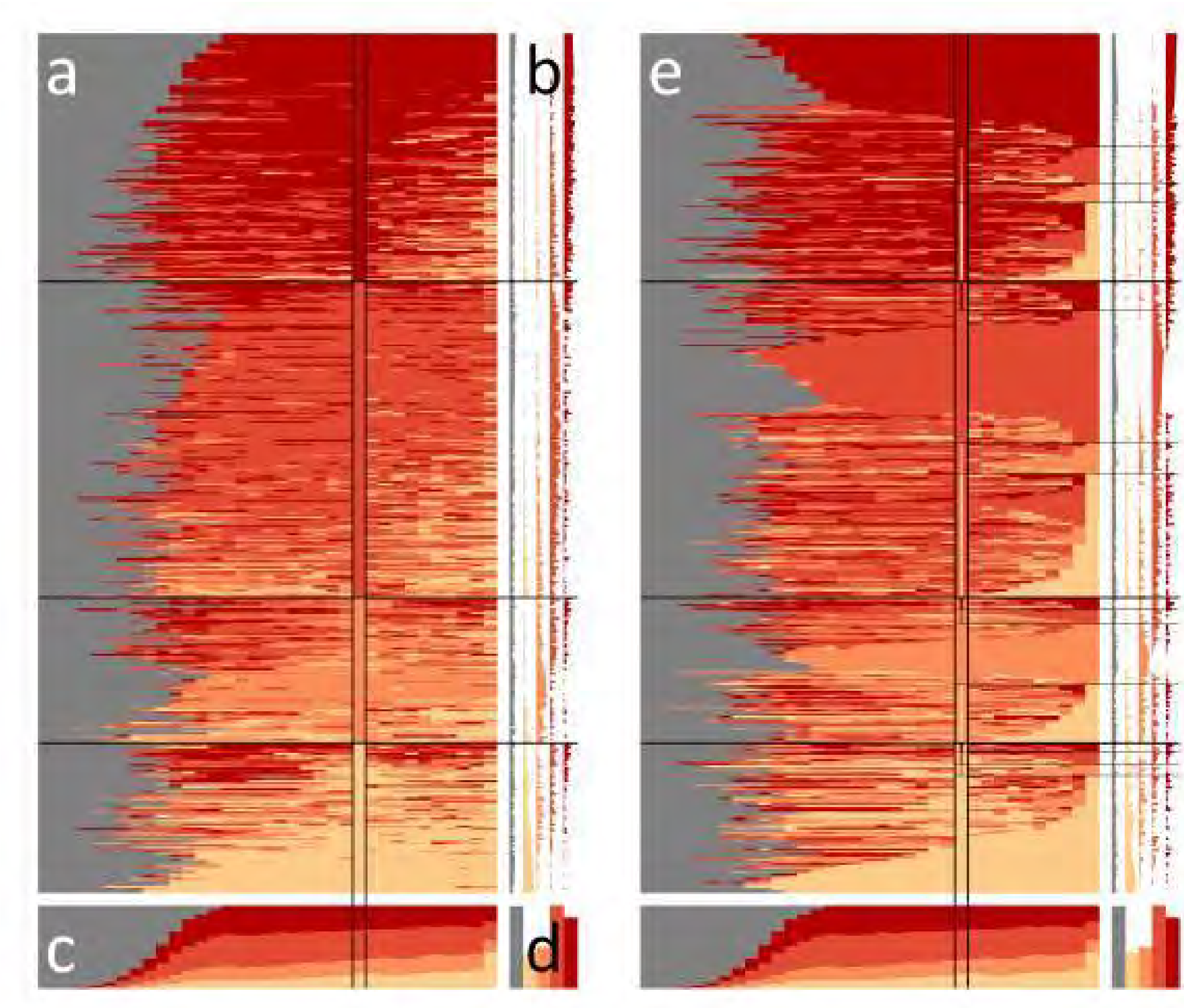


Figure 8: Path map temporal visualization tool

One issue with the parallel-sets based visualization was the inability to show missing data—the most recent measurement is used, even when data is missing for a long period of time. We have therefore made extensions to our path map temporal visualization technique to enable the display of missing data.

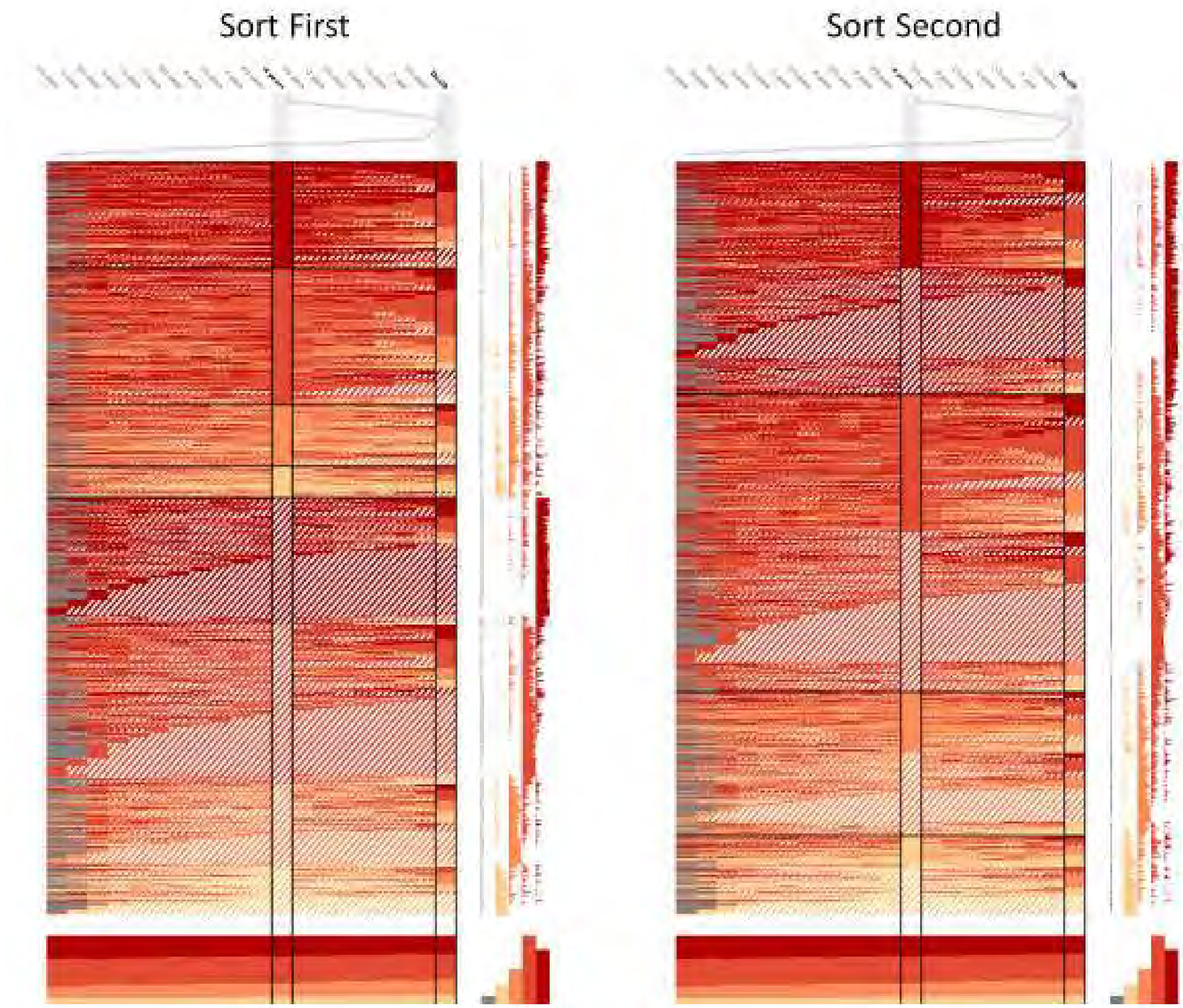


Figure 9: Path maps showing missing data as a striped pattern.

Figure 9 shows HbA1c levels over a 10 year period for 1819 patients with diabetes, and contains a large amount of missing data (no HbA1c level recorded for a given 6 month sampling period). We indicate missing data via a striped pattern. Missing data can either be sorted first (left) to emphasize the total amount of missing data, or second (right), to emphasize the breakdown of missing data for each HbA1c category (Normal: Yellow, Pre-diabetes: Light orange, Diabetes: Dark orange, and Uncontrolled Diabetes: Red). Displaying the missing data in this manner helps us understand how reliable certain conclusions drawn from the visualization might be and has helped us develop rules for the inclusion of patients in our analyses.

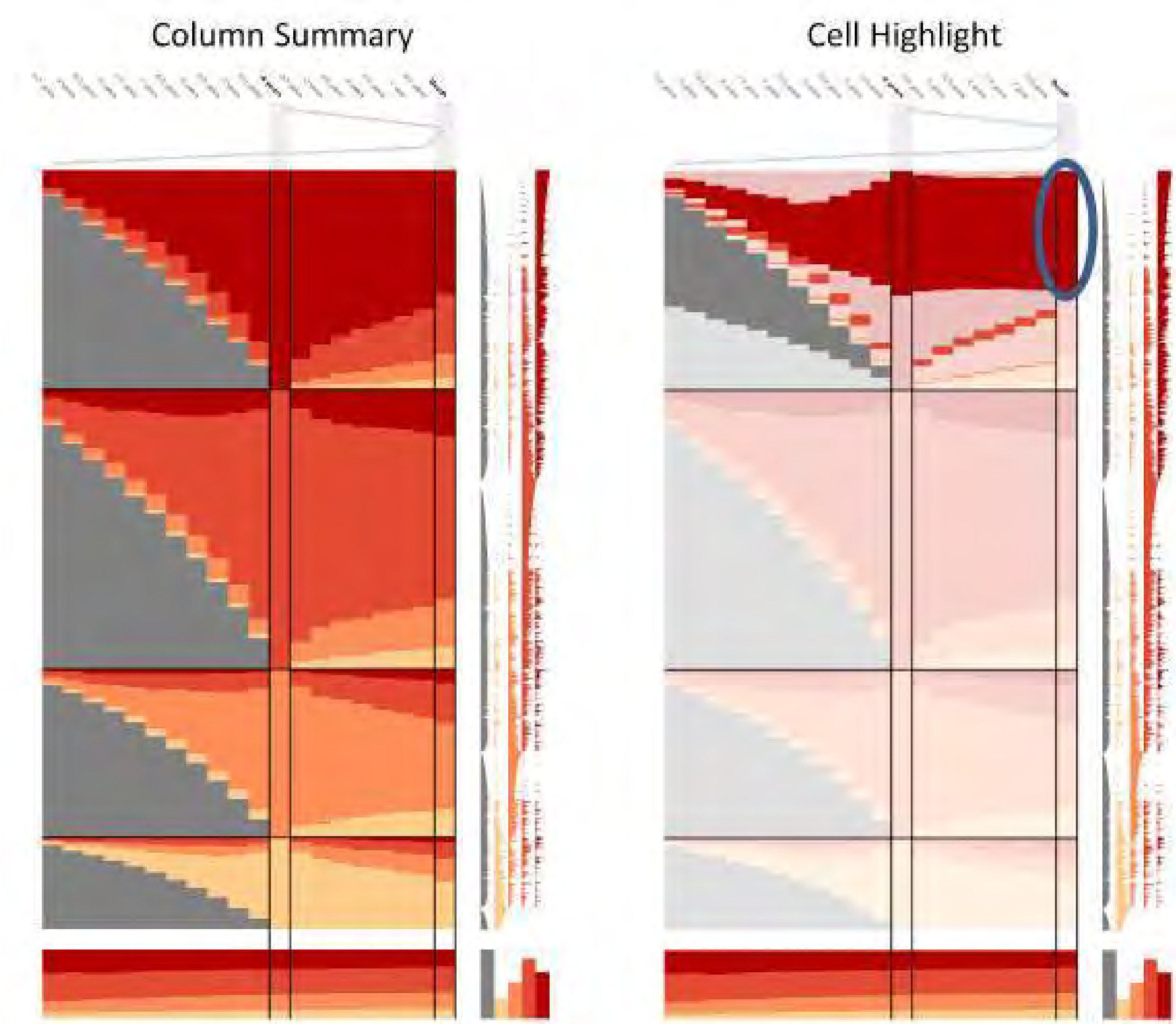


Figure 10: Path maps showing column aggregation for large datasets

The basic path map implementation displays each patient as a single row. This works well for moderately large data sets (hundreds of patients), but overplotting becomes an issue for larger data sets (thousands of patients). Table 10 above is our column aggregation map method, which shows aggregate distributions between user-selected sample columns (left). Highlighting any path map cell shows the distribution of the patients in that cell across the entire path map. We have also implemented two other row-based aggregation techniques. Appendix M provides some more detail on our path map approach, applied to datasets of over 500 and over 3,600 patients.

Although currently we do not incorporate other data, such as age, gender, or race, we plan to investigate this using two methods. One is to include adding columns for each additional variable, similar to the technique used for the parallel sets temporal visualization. The other is to use a supplementary linked view, such as radial coordinates, in which patients could be selected in either view and highlighted in the other.

Visualization of Free Text

Although not directly related to visualization of patient data, we have also experimented with visualizing free text data taken from the visualization in health care literature. Such visualizations could be useful with regards to patient notes. Figure 11 shows a linked view visualization of documents, document clusters, topics, and terms, extracted from a textual analysis of the visualization in health care literature.

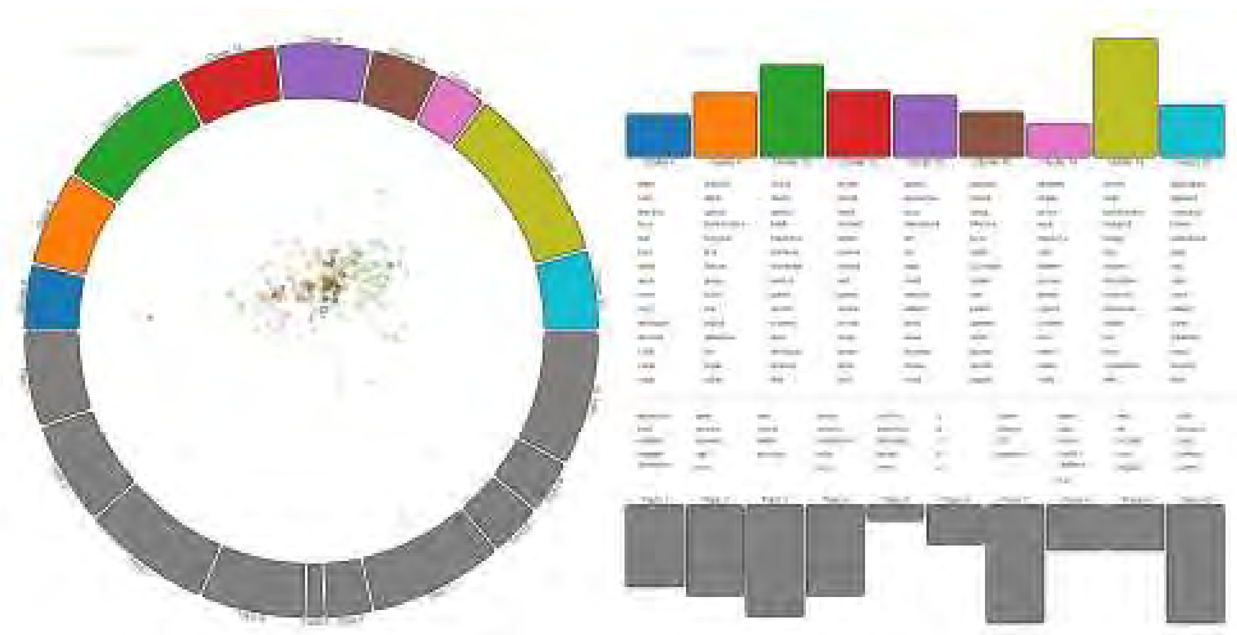


Figure 11: Visualization of visualization in health care literature

The left view shows a scatter plot of documents based on single value decomposition, surrounded by colored clusters and grey topics. The right view shows terms associated with clusters and topics. Selecting any visual element highlights documents, clusters, topics, and terms associated with that element (Figure 12). More information on this approach can be found in Appendices J and N.

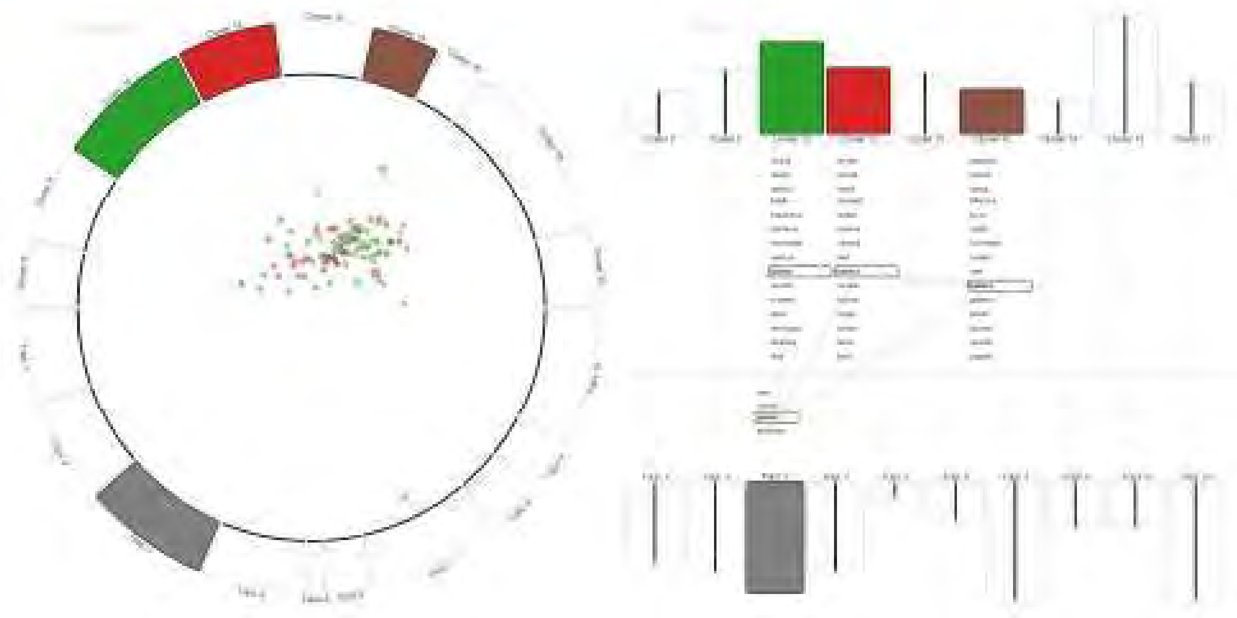


Figure 12. Selecting the term "patient" highlights all clusters and topics with that term, and all documents related to those clusters and topics.

<u>Milestone 7</u>. Develop parallel-coordinates visualization of data resulting from data queries.

(a) Extend parallel-coordinates visualization to include summary statistics per data element and evaluate its ability to reveal significant patterns.

R-based radial coordinates

We developed an initial radial-coordinates visualization, based on parallel-coordinates [6,7] and star plot[8] multivariate visualization techniques in the R programming environment. R is widely used for statistical analysis, including the generation of statistical graphics. We therefore investigated the use of R for developing visualization prototypes, enabling our research team to work together more closely and generate prototypes more rapidly. Additionally, R includes numerous data sets that were useful for prototyping purposes before we had HRPO approval to look at clinical data. To enable some degree of interactivity, we used the rgl library, which provides an interface to hardware-accelerated OpenGL graphics within R.

In our radial coordinates prototype, each data entity (e.g. an individual patient) is represented by lines connecting that data entity's measured value for each axis (e.g. height, weight, race, age, sex, etc.). Our radial coordinates visualization prototype incorporates several enhancements, including:

- Radial axis layout providing a square aspect ratio, which can be beneficial for large numbers of axes
- Axis distribution visualizations based on numeric type (continuous, discrete, categorical)

- Line spreading for integer and categorical data , mitigating the problems of multiple lines collapsing to a single data point for discrete and categorical data, extending the parallel sets[9] visual metaphor to enable visualization of individual data entities, and the incorporation of non-categorical data
- Curved lines to make it easier to visually track along lines
- Automatic axis clustering based on correlations between axes
- Direct visualization of axis correlations via colored arcs connecting axes
- Automatic optional axis flipping based on correlations to minimize line crossings
- Incorporation of pairwise scatterplots for neighboring axes and a central scatterplot based on the first two principal components;
- Interactive line brushing to highlight groups in different colors
- Interactive coloring by axis

Figure 13 shows an example of radial coordinates applied to the same DEDUCE query data from Figure 1, and Figure 14 shows radial coordinates applied to the *mtcars* dataset. Due to the delay in obtaining HRPO approval to look at clinical data, we used various datasets packaged with R, which we also used for developing visualization prototypes. The *mtcars* data includes performance and design data for various automobiles, e.g. miles per gallon, horsepower, and weight. Although not health-related data, this data set includes continuous data, interval data, and categorical data, thus making it a reasonable starting point for working with heterogeneous data. Using this data we were able to look at combinations of the various data types and how relationships can be easily seen across multiple data dimensions.

In this visualization, each axis represents a measured quantity of each automobile model (e.g. miles per gallon), and each automobile model is represented by a curved line connecting its value across the various axes. By highlighting lines in different colors, relationships between individuals and groups of entities can be seen across the data axes. For example, we can see there is a strong relationship between displacement (disp) and weight (wt). This visualization also incorporates per-axis distribution visualizations based on data type, clustering of similar axes, arcs showing correlations between axes, and scatter plots of data entities.

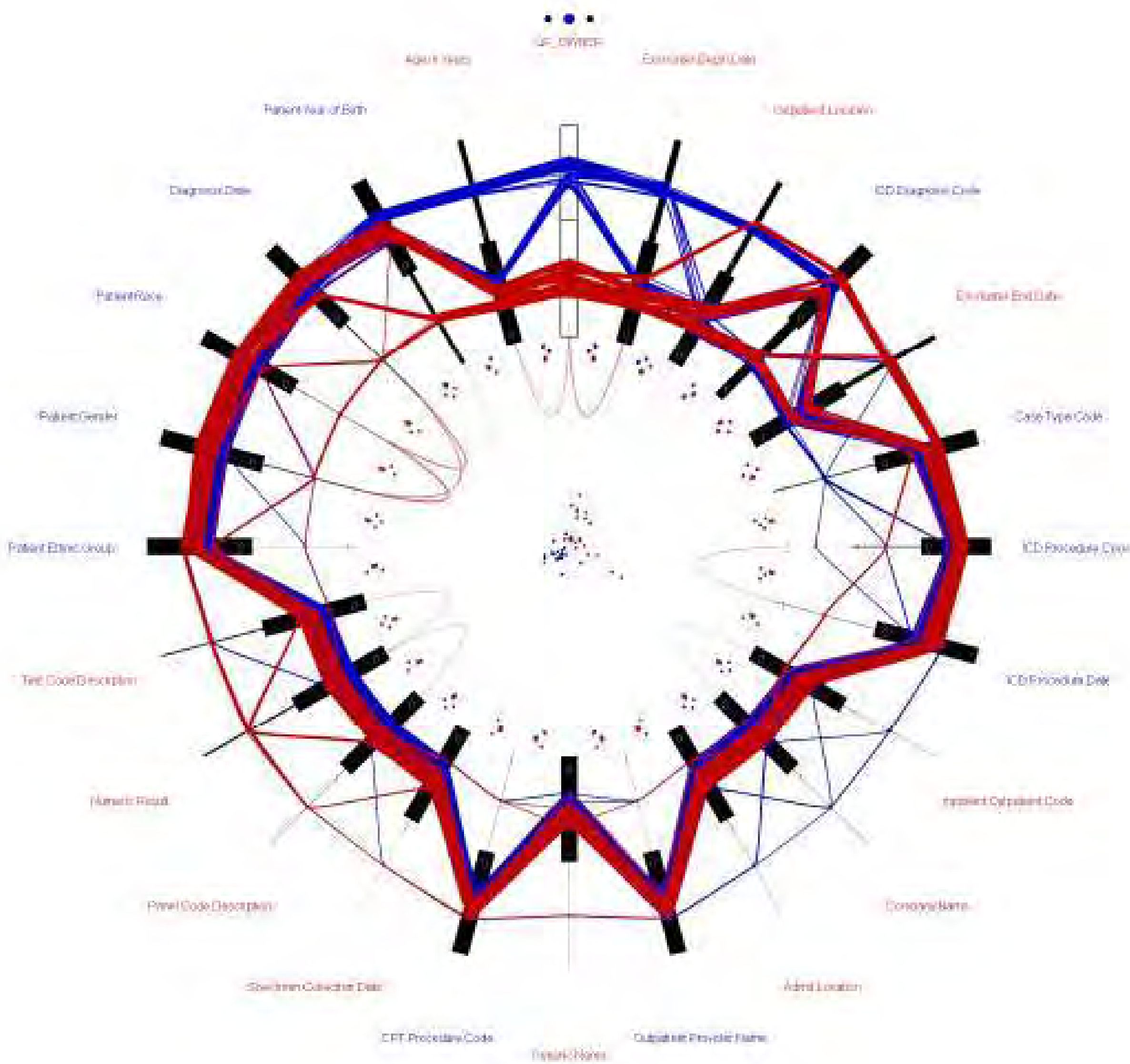


Figure 13. Radial coordinates visualization of DEDUCE queries.

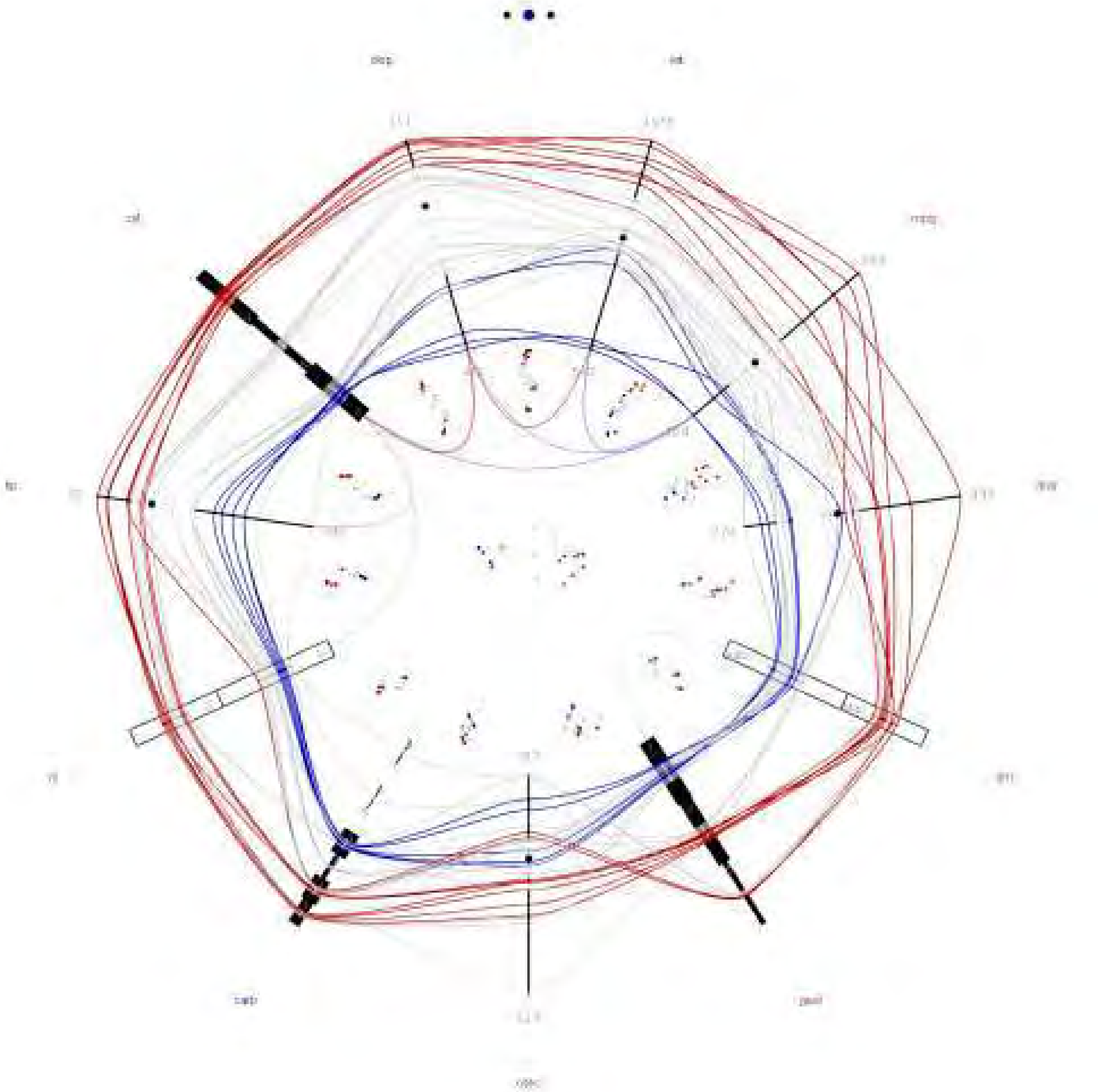


Figure 14. Radial coordinates visualization of *mtcars* data.

Although the R-based radial coordinates prototype proved very useful, especially with respect to the ability to incorporate statistical methods such as correlations and principle components analysis, some drawbacks were determined. These include the lack of support for blended transparency (a common technique for dealing with over plotting in parallel coordinates techniques), and difficulty in incorporating more advanced interactions. We therefore developed a second D3-based radial coordinates visualization.

D3-based radial coordinates

Although our initial plan was to generate prototype visualizations using R and Processing, before creating applications based on the Visualization ToolKit (VTK), we decided instead to use the D3 JavaScript library for development work. This decision was based in part on recommendations from colleagues at the AMIA 2013 VAHC workshop. D3 has quickly become the de facto standard for web-based visualization, and is designed to enable high interactivity.

Our D3-based radial coordinates visualization tool incorporates most features of the R-based prototype. Some preprocessing, such as for principal component analysis (PCA), correlation computation, and correlation-based clustering, is still performed using R. Going beyond the R-based prototype, we concentrated on improving the user interface to the radial coordinates visualization and improving the ability to convey summary statistics of selected groups. These improvements include:

- User control of various parameters, such as opacities, discrete value threshold, and curve properties
- Improved selection capabilities. Users can now easily add to selections, subtract from selections, and select by clicking on various parts of the visualization, such as labels and axis parts
- Per-group ribbons and summary statistic overlays

Figure 15 gives an example of our D3-based radial coordinates visualization applied to published data from the National Health Service (NHS) in the United Kingdom (UK), [10] a data set that became available to us prior to HRPO approval to use clinical data. This data set consists of over 60 million lives spans 2-8 years, with over a billion lines of data. The data are available down to the general practitioner practice or hospital level, and provide information on treatment (in and out patient), pharmacy, demographics, ethnicity, biometrics, and outcomes. It also includes related databases from social care and census; adverse event reporting database from the US Food and Drug Administration; and the VA formulary. The data elements (curved lines) visualized in this example are Primary Care Trust s (PCTs), regional administrative bodies (eliminated in 2013 due to NHS reorganization) responsible for commissioning health services from providers and providing community health services. Here we investigate 26 variables measuring various health and socioeconomic factors for 147 of the 152 PCTs in England (five were removed due to missing data).

In Figure 15, the user has selected a specific London suburb, Harrow, highlighted in red, and has then selected all other London suburbs, highlighted in blue, by changing the selection color and clicking on the *London Suburb* label on the *ONS_Area_Class_Group* axis. In the image on the left, it is possible to compare Harrow to the other London suburbs, but the visualization on the right makes this easier by reducing the opacity of unselected PCTs, and drawing a ribbon outlining the maximum and minimum values for the blue group. The user can then easily see where Harrow is similar to the other London suburbs on aggregate, and where it differs.

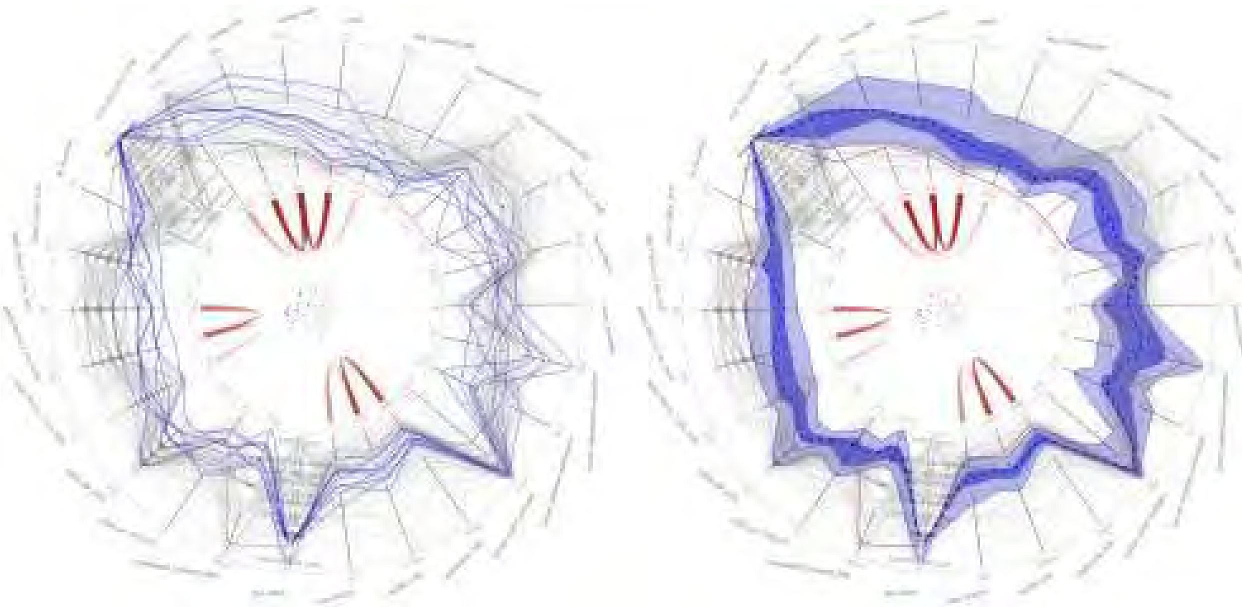


Figure 15. Radial coordinates visualization of NHS Primary.

The largest distinction can be seen in the bottom of the visualization. Harrow has a much higher esophageal cancer rate (Figure 16) than the other London suburbs (although it has the lowest lung, bladder, and colorectal cancer rates across the entire dataset, as visible in the top-right of Figure 15. This ribbon visualization technique is useful for comparing the spread of different user selected groups versus the overall population distribution represented by each axis. Appendix E presents further details on radial coordinates and its application to the NHS dataset.

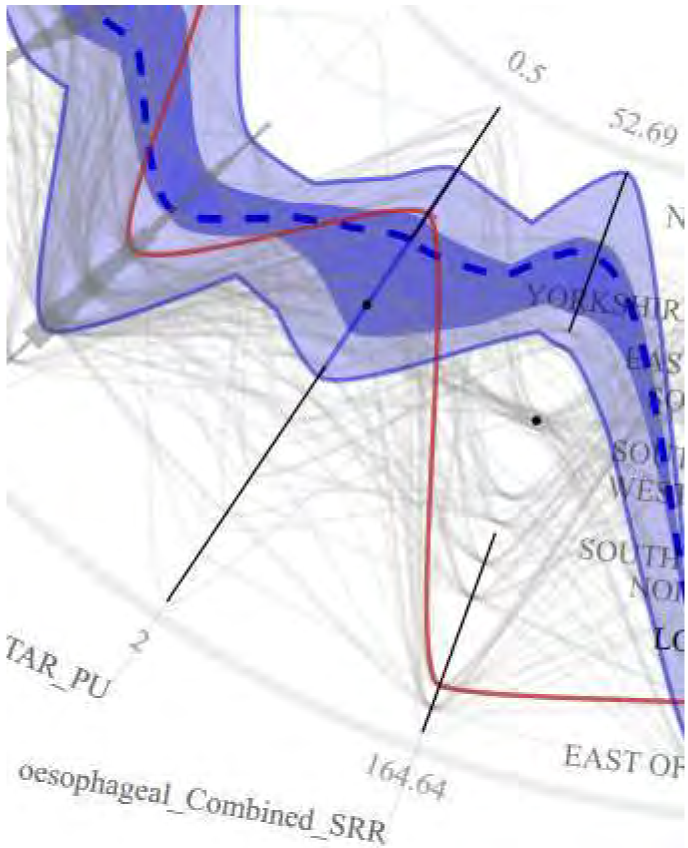


Figure 16: Close up of radial coordinates visualization, highlighted the discrepancy between the Harrow PCT (red) and the other London suburbs (blue) with respect to the *oesophageal_Combined_SRR* axis.

<u>Milestone 8</u>. Add supplemental data dependent linked views of the data based on matrix of visualization techniques.

(a) Experiment with different layout patterns for parallel coordinates.

The traditional parallel axis layout for parallel coordinates has some drawbacks when dealing with large numbers of dimensions, as it can lead to a very wide aspect ratio. We initially experimented with two alternate layout strategies for dimensional axes. The first, a stair-step layout that melded parallel coordinates with 2D pairwise scatterplots (Figure 17).

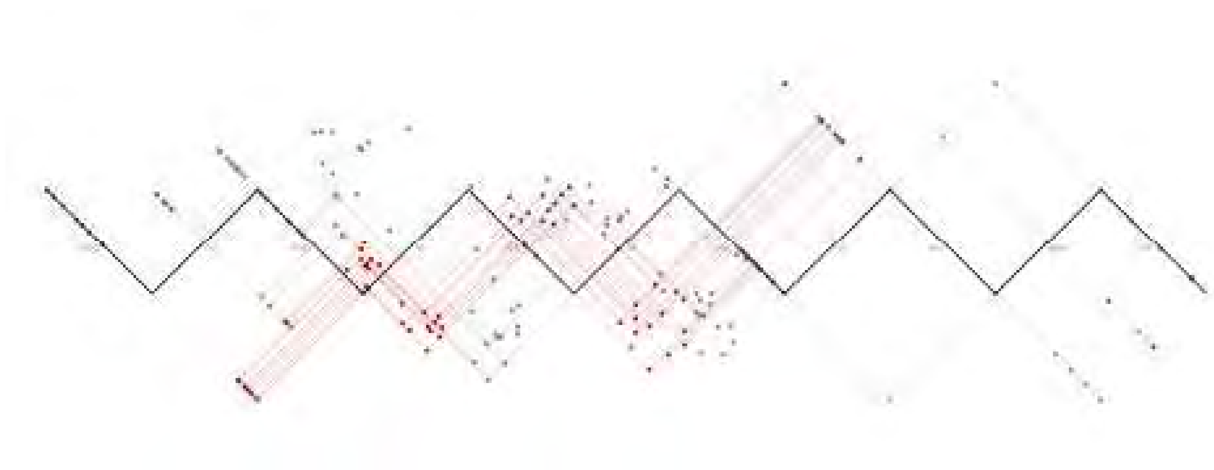


Figure 17. Stair-step axis layout of *mtcars* data.

We quickly determined that this layout was too confusing, and instead concentrated our efforts on using a radial layout. The benefits of the radial layout include a square aspect ratio for the visualization and the ability to utilize the interior area for supplemental visualizations, such as scatterplots and arcs connecting axes to show correlations.

Our radial coordinates visualization makes use of three possible linked views: the radial coordinates view, the PCA scatter plot in the center, and an optional list view of each data element (Figure 18). The user can select curves in the radial coordinates view, points in the scatter plot, or labels in the list view, and see these selections highlighted in the other views. These linked views enable the user to interact with the data in different views to find patterns or individual data entities of interest and visually compare them using distinct visual representations.

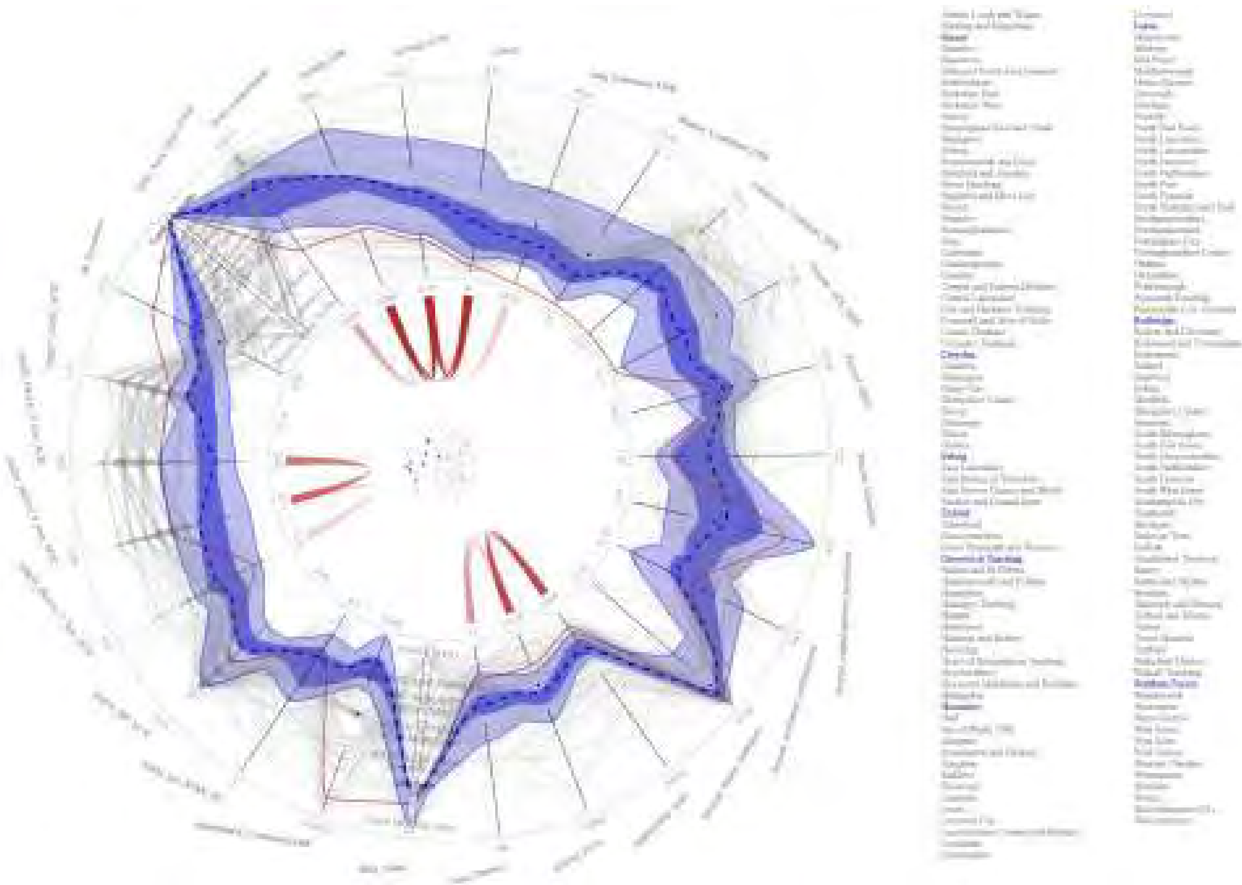


Figure 18. Radial coordinates visualization showing linked views of user-selected data entities.

We also make extensive use of supplemental data dependent linked views in our final visualization of PTSD data, described in Milestone 9.

<u>Milestone 9</u>. Complete testing visualization using PTDS data.

While our radial coordinates visualization was effective for moderately-sized datasets, over-plotting and performance became issues when dealing with larger datasets (thousands of data entities). We therefore developed a new data star visualization to more effectively handle larger datasets. The data star maintains the same radial layout of the radial coordinates visualization, but instead of drawing each individual data element, uses user interaction and stacked bar charts per axis to represent groups of data elements. The data star also uses curved arcs connecting axes to represent the relationship between axes, and includes data-type-dependent linked views to enable drilling-down into data elements of interest.

Data Star Overview

The data star visualization tool, written using the D3 JavaScript library, loads data defined in a custom JSON meta data file, which defines a *data collection* comprising multiple *datasets* (separate CSV files), which in turn comprise multiple *data elements* (individual columns, or groups of columns in the case of hierarchical data). The data in Figure 10 include 2,127 patients with a primary diagnosis of PTSD in data from Duke's EHR. Each data element is assigned a data type (*normative*, *categorical*, or *hierarchical*), which determines the type of visualization used for that data element in the *focus view* described below.

Figure 19. Data Star Visualization

Our data star visualization tool, shown above, comprises five linked panes:

1. *Data overview and controls*. Provides information about the loaded data collection, including number of patients, along with axis threshold and group selection controls.
2. *Data element selection*. A list of available data elements, grouped by data set.
3. *Data star visualization*. Overview visualization of the currently selected data elements, grouped by dataset.
4. *Focus view*. Data-type-dependent visualization of a user-selected data element (in this case a hierarchical icicle plot of ICD-9 codes).
5. *Group lists*. Lists of the user-selected patient IDs in group 1 and group 2 (currently empty).

These components are discussed in more detail below.

Data Overview and Controls

The data overview and controls pane consists of a label displaying the name of the data collection (PTSD data in this example), and the number of patients in the collection. The axis threshold slider enables filtering of axes in the data star by the number of patients with the data value represented by that axis, which can help remove visual clutter, such as with the *3-Digit ZIP* data element highlighted below.

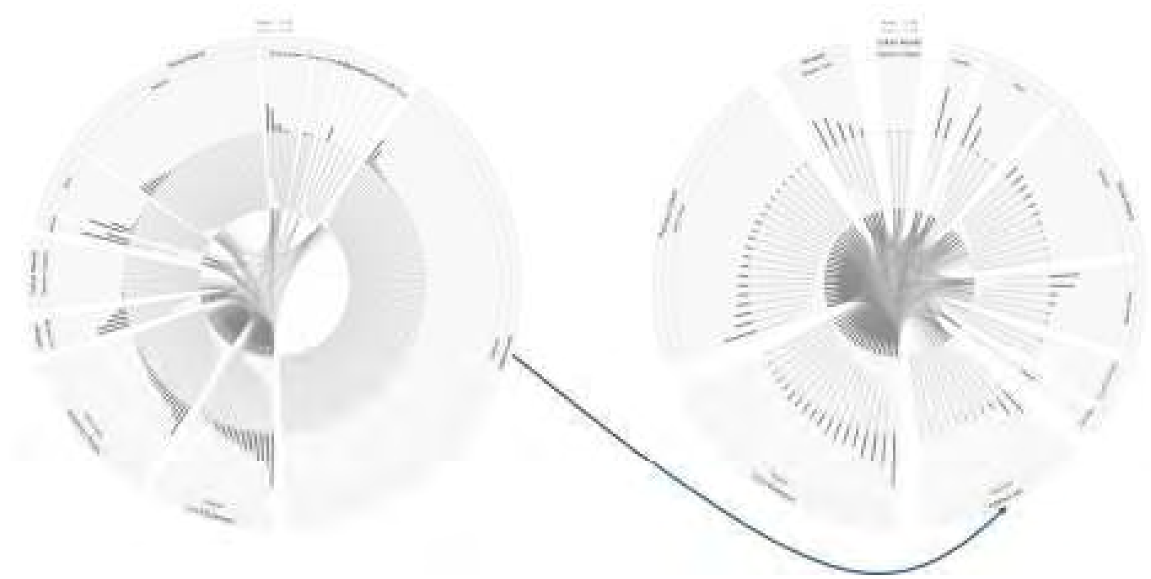


Figure 20. Using the axis threshold slider to reduce the number of axes with small values

Group selection controls enable choosing the current group (group 1 or group 2) and the group operation (assign, union, intersection, or difference) to be applied when the user creates groups of users by interacting with the data star and focus visualizations, described below.

Data Element Selection

The data element selection pane displays a list of available data elements, grouped by data set. The user can turn on and off display of any data element by clicking on it in the list. Below on the right the user has deselected a number of data elements in the *Social History* dataset, as well as *3-Digit ZIP* in the *Address Info* dataset. This data element selection ability enables the user to hone in on data elements of interest.

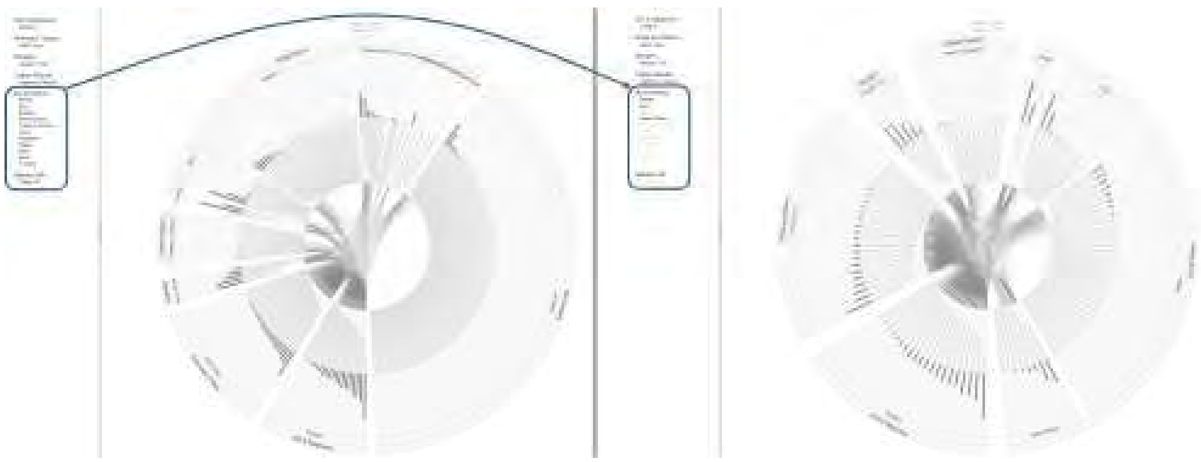


Figure 21. Using the data element selection pane to filter data elements of interest

Data Star Visualization

The data star visualization pane contains the data star overview visualization. The outer ring is divided into sections by dataset, and then each dataset is divided into one or more data elements. For example, the image below highlights the *Social History* dataset, which contains the *Gender*, *Race*, *Religion*, *Marital Status*, *Types of Alcohol*, *Chew*, *Cigarettes*, *Cigars*, and *IV Drugs* data elements.

Figure 22. Data star visualization, with Social History dataset highlighted

Each data element contains an axis for each value contained in the data element (for hierarchical data elements, only the top-level data values are displayed). Each axis consists of a bar representing the entire population. The darker portion of the bar drawn above the dividing circle shows the proportion of the population with that value, and the lighter portion below the dividing circle the proportion without that value. Due to the large number of axes, the label for each axis is only displayed when highlighted by the user, along with the number and percentage of patients with that value. The central portion of the data star displays arcs representing the co-occurrence frequency between each data element value. Highlighting a single data value will show the co-occurrence frequency arcs for just that value, as shown below.

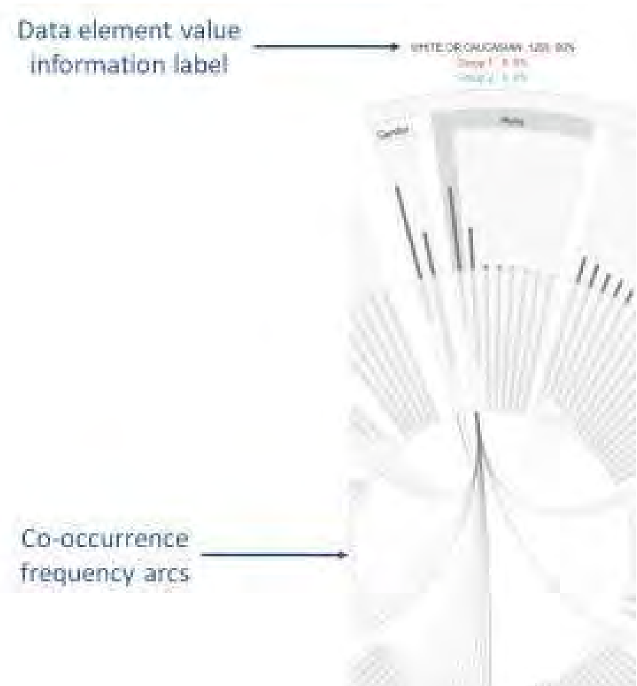


Figure 23. Co-occurrence frequency arcs highlighting a single data value

The user can create two groups of patients by combining data element values in different ways using the current group and the group operation controls and selecting different data elements. All axes are then partitioned based on the proportion in group 1 (red), group 2 (blue), and the intersection (purple). The group labels above the data star indicate the current data elements combinations used for each group, along with the number and percentage of patients. The example below shows all patients prescribed *Gastrointestinal Drugs* in red, all patients with *Diseases of the Digestive System* in blue, and the overlap in purple. Providing user-defined patient groups based on data element values in this manner enables the user to see how these patients are distributed across all currently visible data element values.

Figure 24. Distribution of selected patient groups across all data element values

Focus View

By selecting any data element in the data star, a data-type-dependent visualization is displayed in the focus view pane. Currently two such visualizations are possible, an icicle plot for hierarchical data elements, and a parallel sets visualization for multiple normative and categorical data elements.

Icicle plot

The icicle plot is a space-filling technique that shows multiple levels of a hierarchy via rectangles with the height of leaf nodes proportional to the number of patients with that data value, and the height ancestor nodes sized to fit child nodes. In our implementation we also adjust the greyscale color of each rectangle to represent the number of patients with that data value, based on the highest-frequency leaf node. This feature helps differentiate between many nodes. We also enable mapping patient frequency to rectangle width, to further accentuate differences between nodes. The image below shows an icicle plot of all *ICD-9 Diagnosis* codes present in our PTSD data collection, with the *Anxiety state, unspecified* diagnosis code (and its ancestor subcategory and category) highlighted.

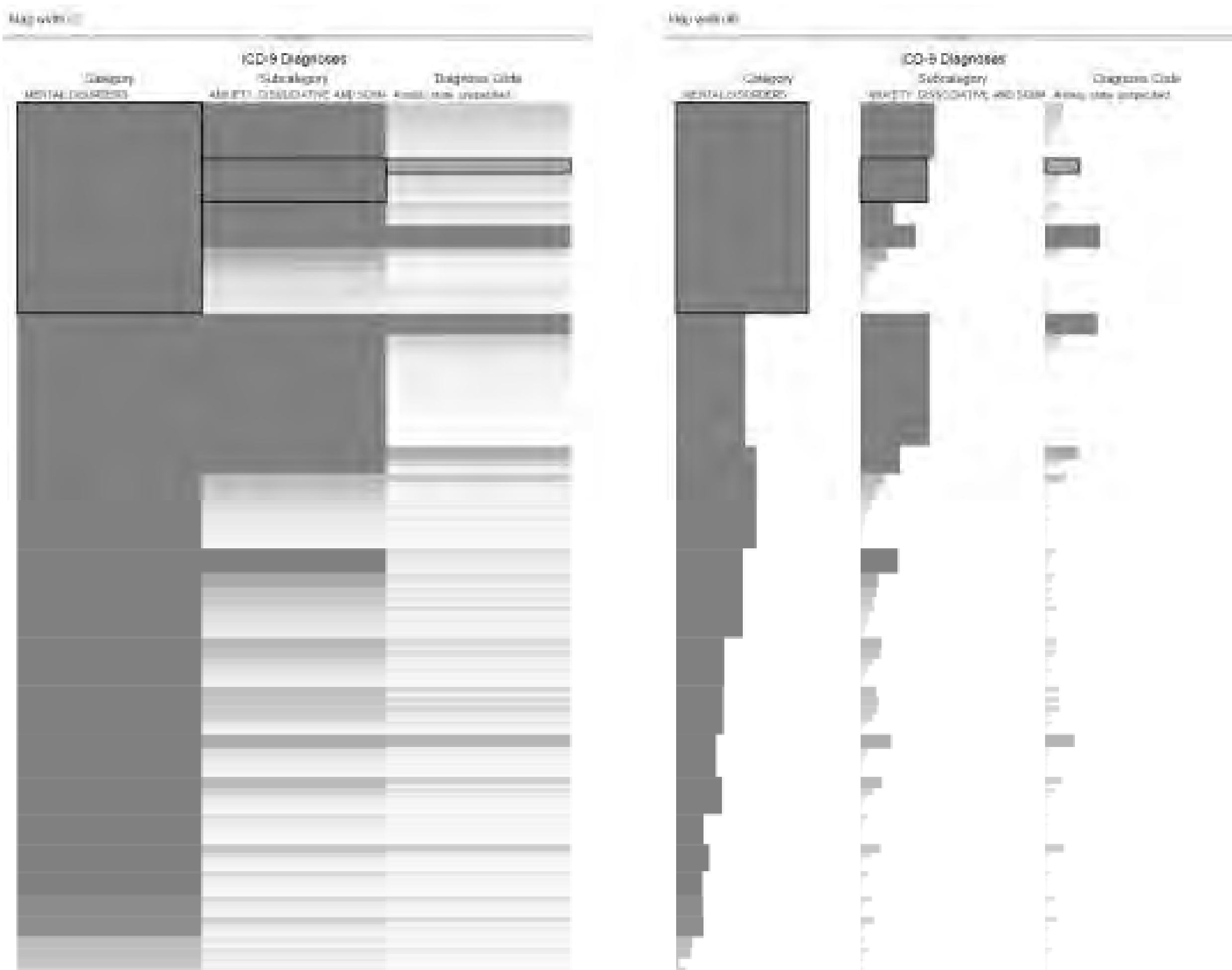


Figure 25. Focus view pane with icicle plot visualizations of hierarchical ICD-9 diagnosis codes. The user can select whether to map the width (right) or not (left).

Users can select any data value in the icicle plot to define groups, as with the data star. Groups created in the icicle plot are shown in the data star, and vice versa. In Figure 26 below the user is displaying the *AFHS Class* and *Therapeutic Class* of patient *Pharmacy Orders* in the icicle plot, and has created groups with all patients prescribed *Psychotherapeutic Drugs* in red and *Narcotics C-IV* in blue (overlap in purple). Looking at the breakdown of these colors across all therapeutic classes indicates that many drugs seem to overlap mostly with one or the other, such as *Antidepressants* (highlighted with arrow), which contains mostly blue and very little red, indicating a common co-occurrence with *Narcotics C-IV*, but not *Psychotherapeutic Drugs*.

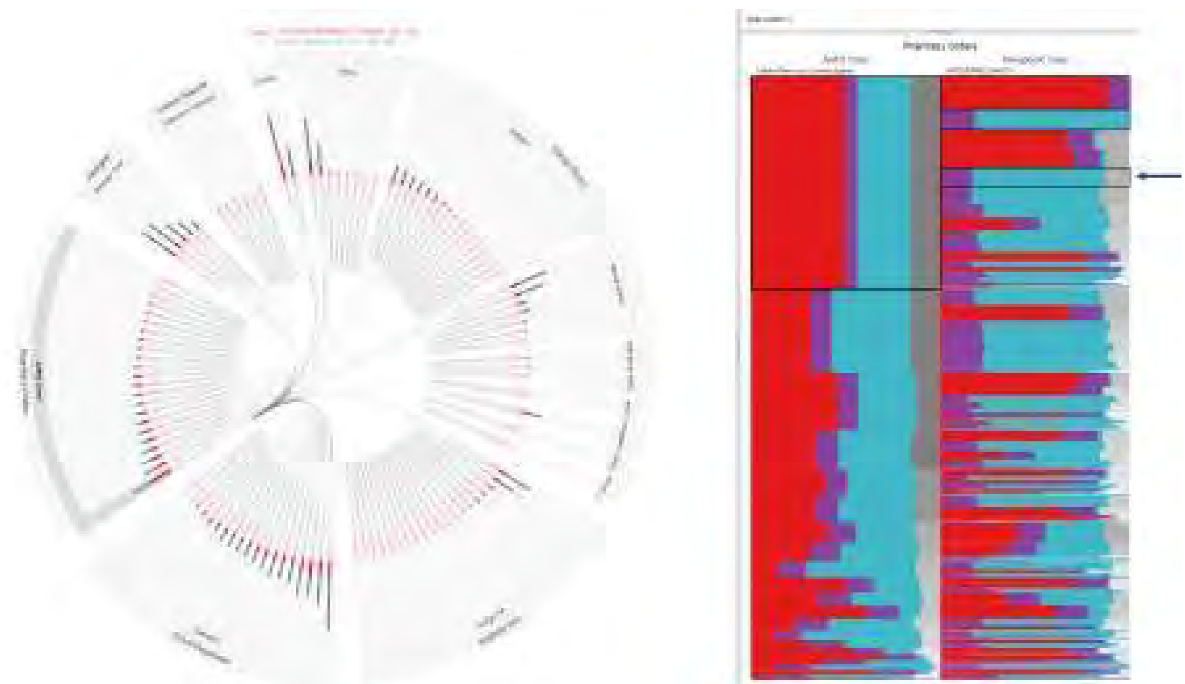


Figure 26. Highlighting different pharmacy orders in the icicle plot focus view, with linked selection in the data star visualization.

Parallel Sets

If the user select a normative or categorical data element, that data element is added as a horizontal axis in a parallel sets visualization, which shows the frequency of combinations of data values via the width of curved paths through those data values at each data element. Data element axes can be interactively rearranged vertically, and data values can be rearranged horizontally per axis. The proportion of each data value combination belonging to one of the currently selected groups is conveyed using coloring by whichever group is more frequent, with the intensity of the color indicating the frequency.

In Figure 27 the user has added *Patient Gender* and *Cigarettes Indicator* to the parallel sets visualization, and added patients with *Drug Ingredient* allergies to group 1 (group 2 is empty). The curved path representing female smokers has a more intense red color, indicating a higher proportion of drug ingredient allergies in that group. Selections of data value combinations, such as female smokers, can also be made via selection in the parallel sets visualization (Figure 27, right), which will be reflected in the data star, as with the icicle plot.

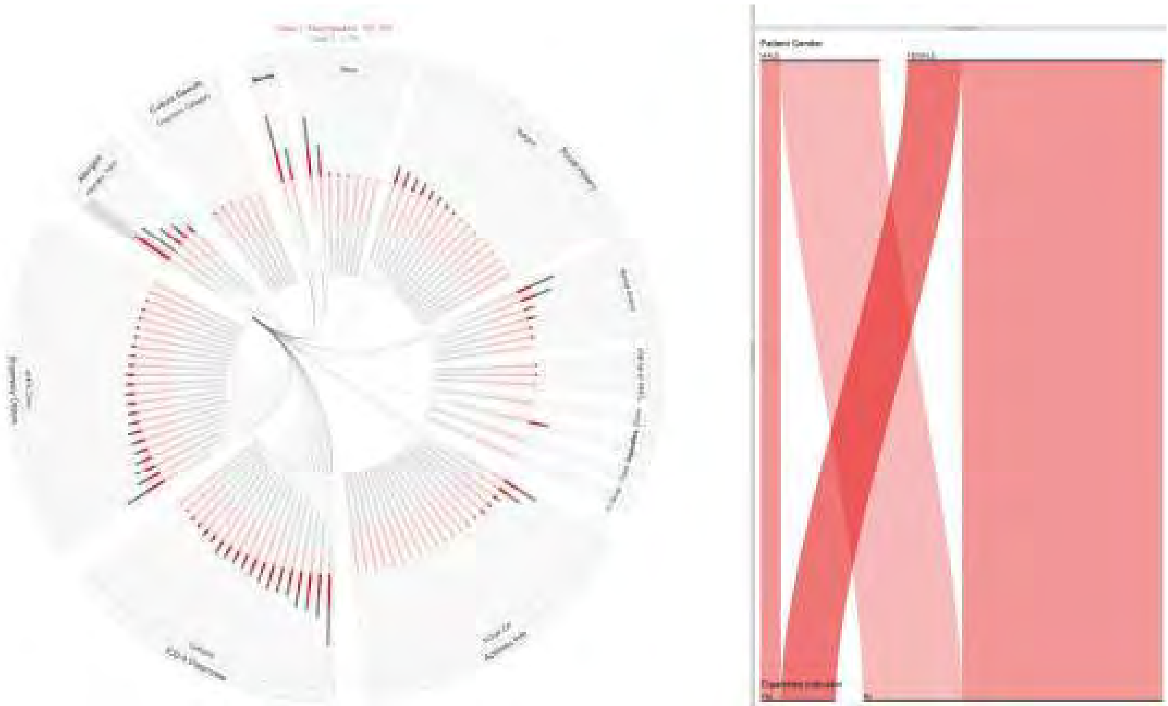


Figure 27. Focus view pane on the right using parallel sets visualization of categorical data

Group lists

There is a pane at the bottom of the visualization displays all (de-identified) patient IDs in the currently selected groups. This feature makes it easy to extract selected ids for offline analysis.

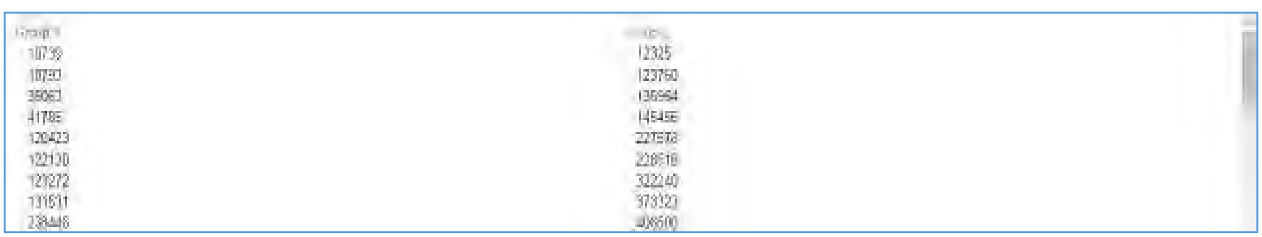


Figure 28. Group lists pane

The Data Star Visualization closely resembles the visualization we originally proposed in 2012. As our research accomplishments in Section 4 describe, we have made significant advancements in this novel approach to visualization big data from electronic health records. The ability to drill down into different types of data and visualize groups of data can be done, as we have shown. We have not yet determined a realistic method to examine individual data from our population-focused data star, but its value to examine differences in populations (e.g. Army versus Marines, or deployed versus not deployed) has great potential.

4. KEY RESEARCH ACCOMPLISHMENTS

We are excited to be involved interactive visualization of big data that is just beginning to become recognized for its potential. The following completed goals and objectives are research contributions that we believe will have significance to this field,

- One of our most significant accomplishment is the work looking at the temporal course of relevant data elements for patients with the same diagnosis (ICD-9 codes). Aligning the patients at some point in time (such as death) and examining the patterns presented enabled us to identify subpopulations of the disease. This finding is critical in precision medicine leading to patient specific recommendations for treatment as well as predicting outcomes. There is an economic factor in knowing how to optimize treatment to be most effective.
- We visualizing unique types of data, e.g. environmental data and population health. (BT data), providing the evidence that knowledge otherwise not discoverable can be discovered using unique visualization techniques.
- Visualization is an effective way to look at big data. More than one visualization algorithm is important because different techniques reveal different understandings.
- Data visualization is an excellent way to mine the literature, which is translatable to text data in EHRs.
- It is difficult to look at all types of data at once. A more effective use of visualization is to look at subsets of data (for example comorbidities) and visually display what abnormalities are a result of a particular disease.
- The complexity of the presentation of data requires the use of additive measures of points of interest, which is difficult to understand without the ability to interactively highlight certain types of data to see various correlations and highlight the relationships among the data elements.
- Users must recognize that data contained in the EHR is random in the sense that it occurs (now) only when a patient encounter occurs. When analyzing data across many patients, the time interval between data entry varies widely. What happens between data points is significant in the analysis. We discovered the value of filling in the gaps with imputed values based on an equivalent test of statistically derived.
- There are multiple ways of creating synthetic data sets. Their value is limited, depending on the method used for creation of the data set.
- Results support our hypothesis: data visualization is more effective than traditional methods of data exploration, and the type of visualization is highly dependent on the types of data, nature of the query, and what someone is trying to learn from the data.

5. CONCLUSIONS

Current emphases in health and health care include Big Data, Learning Health Systems, and Precision Medicine. Data visualization is an important tool in each of these. We have shown that interactive visualization of large data sets leads to better understanding of what is in the data. Although we are likely early adapters to understanding the value of this approach, our research supports our hypothesis that data visualization is more effective than traditional methods of data exploration, and that the type of visualization is highly dependent on the types of data and what one is trying to learn from the data. We did not have an opportunity to use AHLTA data as planned and therefore cannot generalize what we learned using Duke data to this population and its military significance. Exploring the EHR data of military personnel using the data star visualization methods we have described, however, can offer new insights into the health and resilience of this population.

Future Plans

Using visualization, we explored primarily two diseases: type 2 diabetes mellitus and PTSD. Our work with PTSD data allowed us to develop a methodology to drill down into various data elements for several groups. Our next step will be to increase the number of data elements that can be visualized using the data star visualization, and to explore additional visualizations that represent data from specific data elements (such as parallel sets or icicle plots described in Milestone 9). We also want to focus on usability for the various visualizations and conduct a usability analysis with the data star.

In further research we plan to explore other diseases including hypertension, kidney disease, congestive heart disease, breast cancer, and multiple sclerosis. We would like to have an opportunity to apply the tools we have developed applying them to DoD data to see if the data star visualization findings are generalizable to that data.

By adding specific data elements as a component of the temporal analysis, we can learn the value of that data element to the identification of subpopulations of diseases. We did not have time to investigate extensions of temporal trends to predict the course of a disease and to extending backwards to better understand the first indications of a disease. This is a step our future research will address.

6. PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS

Please see Appendix for copies of all publications, abstracts, and/or presentations.

a. Publications and abstracts.

(1) Lay Press

Nothing to report.

(2) Peer-Reviewed Scientific Journals

- West, V. L., Borland, D., & Hammond, W. E. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, *22*(2), 330-339. (2015) DOI: 10.1136/amiajnl-2014-002955. PMID: 25336597

- West, V.L., Borland, D., & Herhold, L. Visual Presentation of Data: A Qualitative Study of Acceptability and Usability. Pending review.
- McPeek Hinz, E., Spratt, S., Borland, D., Herhold, L., West, V.L., Hammond, W.E., Akushevich, I. Interactive Visualization of Temporal Diabetes Data. Pending review.

(3) Invited Articles

Nothing to report.

(4) Abstracts

- Borland, D, West, V., Hammond, WE. Demonstration of Visualization of EHR and Health Related Data for Information Discovery. Proceedings of the 2013 Workshop on Visual Analytics in Healthcare: Washington, DC, 96. (16 November 2013). http://www.visualanalyticshealthcare.org/docs/VAHC2013_proceedings_LowRes.pdf
- Shah, H., Borland, D., McPeek Hinz, E., West, V.L., & Hammond, W. E. Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 53-58. (15 November 2014) https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf

b. **Presentations** (An asterisk (*) is a presentation that produced a manuscript.)

- (*) West, V., Borland, D, and Hammond, WE. Visualization of EHR and Health Related Data for Information Discovery. Proceedings of the 2013 Workshop on Visual Analytics in Healthcare: Washington, DC, 33-36. (2013) http://www.visualanalyticshealthcare.org/docs/VAHC2013_proceedings_LowRes.pdf
- Borland, D. Visualization of Health Informatics Data. Duke Visualization Friday Forum: Duke University, Durham, NC. (4 April 2014) http://vis.duke.edu/FridayForum/?semester=Spring_2014#2014-04-04
- (*) Borland D., West, V., and Hammond, WE. Multivariate visualization of system-wide National Health Service data using radial coordinates. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 53-58. (15 November 2014) https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf
- (*) McPeek Hinz, E., Borland D, Shah, H, West, V., and Hammond, WE. Temporal visualization of diabetes mellitus via hemoglobin A1c levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 17-22. (15 November 2014) https://dl.dropboxusercontent.com/u/4724665/VAHC2014_proceedings.pdf
- Ganapathiraju, M., Borland, D., West, V., Hammond, WE. Exploring Novel Visualizations of Electronic Health Record Data. Proceedings of AMIA 2014 Annual Conference, High School Scholars: Building New Paths to Biomedical Informatics Education. Washington, DC. (17 November 2014)
- (*)West V. Innovative Information Visualization of Electronic Health Record Data: A Systematic Review. JAMIA Monthly Book Club: Webinar. (5 February 2015)

- West VL; Borland D, West D, Hammond WE. Visualization of the Health Care Visualization Literature. AMIA Joint Summits 2015: San Francisco, CA. (26-Mar-2015)

- West, V., Borland, D. Visualization of Health Care Data. Duke Clinical Research Institute: Durham, NC. (26 August 2015).

- West V. Visualization of Health Care Data. North Carolina Healthcare Information and Communications Alliance (NCHICA) 21st Annual Conference: Pinehurst, NC. (14 September 2015)

- (*) Borland, D., McPeek Hinz, E., Herhold, L.A., West, V.L., Hammond, W.E. Path Maps: Visualization of Trajectories in Large-Scale Temporal Data. IEEE Vis 2015: Chicago, IL. (28 October 2015) https://vimeo.com/136251962

- (*) West, V.L., Borland, D., West, D., Hammond, W.E. An Evaluation of Machine Learning Methods and Visualization of Results to Characterize Large Healthcare Document Collections. 2015 Annual Meeting of the Decision Sciences Institute: Seattle, WA. (23 November 2015)

7. INVENTIONS, PATENTS AND LICENSES

Nothing to report.

8. REPORTABLE OUTCOMES

- We implemented a prototype radial coordinates visualization tool in the R programming environment, and a prototype force-directed node and link visualization tool in the Processing programming environment to use with large data sets.
- A systematic literature review was conducted to identify how visualization is used with health care data was completed with a manuscript of results published in JAMIA.[3] The paper was acknowledged by Caban and Gotz who state that the review paper "is a great resource for readers who are interested in understanding what has been done in applying visual analytics in healthcare settings"[11] and by Dr. Ben Shneiderman, one of the leading researchers in the field, in an email to the authors as a "thorough and helpful review paper on EHR visualization", and it has been cited in several manuscripts since its publication this year.
- We refined radial coordinates visualization, implementing a new, more interactive version using the D3 JavaScript library, and applied it to both Primary Care Trust (PCT) and practice-level data from British Telecom Cloud service (BT), data that comes from published health data from the UK National Health Service. Visualization of lung cancer rates among PCTs discovered possible relationships between lung cancer rates, socioeconomic factors, and regional classification. A comparison of London-based PCTs revealed a potentially interesting PCT with much higher esophageal cancer rates than

other similar PCTs. Visualizing medical problem prevalence among over 1,500 London practices showed two practices that have much higher rates of medical problems. Results were presented at VACH 2013.[12]

- We developed a parallel-sets based temporal visualization tool and investigated its use as applied to HbA1c levels from diabetes patient data extracted from a DEDUCE query. Through use of this tool we have discovered that in the 10 years before death there is a consolidating trend to improved glucose control across all diabetes control categories from uncontrolled to normal. By including category temporal transitions this visualization also illustrates the complexity of the underlying data, with many trajectories exhibiting a large degree of variation in HbA1c categorization over time. Results were shared at VAHC 2014[13] and a manuscript is pending review.
- We incorporated axis clustering, correlation chords, and a PCA scatterplot in D3-based radial coordinates visualization.
- We added automatic categorical axis value reordering to radial coordinates visualization.
- We developed a novel path map temporal visualization tool using DEDUCE diabetes data. Results were presented at IEEE Vis 2015.[14]
- We made extensions to our path map temporal visualization technique to enable the display of missing data and aggregated information to handle larger data sets. This is described in Appendix M.
- We developed D3-based linked views for visualizing the visualization in health care literature.
- We completed proof-of-concept connecting an R Shiny server from JavaScript/D3. This has the potential for being able to do various statistical calculations in R with that data then passed to JavaScript/D3 for visualization. A Python back-end may also be used with various statistical packages for Python. Future research will allow us to explore this proof-of-concept further.
- We developed a new health data star multivariate visualization technique and tested it using de-identified data from Duke patients with PTSD. This is describe in Milestone 9.
- Outcomes from our research were shared through 14 different publications, presentations, posters and two manuscripts that are pending review.

9. OTHER ACHIEVEMENTS

We submitted a grant application to the National Institute of Biomedical Imaging and Bioengineering for a funding opportunity for Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science (RO1). Unfortunately, our application did not score high enough for funding. We are currently revising the application to address the helpful suggestions of the review panel and plan to submit it in February. The application is based on the research completed as described in this Final Report.

We have also submitted an application and are waiting for results for a CTSA-sponsored institutional grant that is designed for teams composed of researchers from Duke University and the University of North Carolina at Chapel Hill. This is a small, one-year project that also builds on the research completed as described.

We are hopeful that additional funding from the U.S. Army Medical Research and Materiel Command will become available and that we will have an opportunity to continue the research from the point we stopped. We have validated our hypothesis that data visualization is more effective than traditional methods of data exploration, and that visualization of big data to see what is in the data will lead to knowledge discovery. As we have moved closer to developing an effective interface that allows users to interactively explore big data, we believe its application for military purposes has significant promise. Being able to detect causal relationships between various data sets, particularly from personnel in different branches of service, military ranks and job specialties, ages, and geographical deployment areas has the potential to lead to not only improved health care and resiliency for military personnel, but could also assist the DoD in strategic decisions related to personnel, and save millions of dollars in health care costs for not only the Department of Defense, but for society. Our work for this project has used Duke EHR data; we hope that the research might continue another data source to validate our discoveries thus far.

10. REFERENCES

1. Horvath, M. M., Winfield, S., Evans, S., Slopek, S., Shang, H., & Ferranti, J. (2011). The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. Journal of biomedical informatics,44(2), 266-276.
2. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.*Annals of internal medicine*, *151*(4), 264-269.
3. West, V. L., Borland, D., & Hammond, W. E.(2015) Innovative information visualization of electronic health record data: a systematic review. Journal of the American Medical Informatics Association, 22(2), 330-339.
4. Kosara R, Bendix F, Hauser H. (2006) Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics,12(4):558-568.
5. Wilkinson, L. (1979), Permuting a matrix to a simple pattern," in Proceedings of the Statistical Computing Section of the American Statistical Association, Washington, DC: The American Statistical Association, 409-412.
6. d'Ocagne M. (1885) Coordonnees parallels et axiale. Gautier-Villars, Paris.
7. IInselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, *1*(2), 69-91.
8. Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983) Graphical methods for data analysis. Wadsworth, Belmont, CA.

9. Bendix F, Kosara R, Hauser H. (2006) Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006; 12(4):133-140.

10. National Health Service Data. BT Group: London, UK, 2013.

11. Caban, J. J., & Gotz, D. (2015). Visual analytics in healthcare–opportunities and research challenges. *Journal of the American Medical Informatics Association*, *22*(2), 260-262.

12. West, V., Borland, D, and Hammond, WE. (2013) Visualization of EHR and Health Related Data for Information Discovery. Proceedings of the 2013 Workshop on Visual Analytics in Healthcare: Washington, DC, 33-36.

13. McPeek Hinz, E., Borland D, Shah, H, West, V., and Hammond, WE. (2014) Temporal visualization of diabetes mellitus via hemoglobin A1c levels. Proceedings of the 2014 Workshop on Visual Analytics in Healthcare: Washington, DC, 17-22.

14. Borland, D., McPeek Hinz, E., Herhold, L.A., West, V.L., Hammond, W.E. (2015) Path Maps: Visualization of Trajectories in Large-Scale Temporal Data. IEEE Vis 2015: Chicago, IL.

11. APPENDICES

Appendix A

Visualization of EHR and Health Related Data for Information Discovery

Vivian West[1] David Borland[2] W. Ed Hammond[1]
[1]Duke Center for Health Informatics, Duke University
[2] Renaissance Computing Institute, The University of North Carolina at Chapel Hill

Abstract

In this paper we describe research we are conducting in response to a Program Announcement solicited by the Assistant Secretary of Defense for Health Affairs, Defense Health Program. The amount of information in Electronic Health Record (EHR) systems is growing rapidly with the inclusion of disparate forms of data from a number of new sources, i.e. genomics and imaging data. EHR systems will continue to grow as more healthcare data is digitized. As data in EHRs grows, there is increasing interest in understanding what information and knowledge these large data sets represent.

Data visualization techniques offer an opportunity to explore and understand large data through novel approaches. Our research seeks to visualize health care data from electronic health records (EHR) and other health related data. Our approach is informed by retrospective data queries using DEDUCE, a query tool developed at Duke University.

Keywords: Electronic health records, health related data, information visualization

Introduction

Visualization of genomic data is used to understand data structures. Geospatial applications have revealed patterns related to risk factors in environmental health,[1,2] and visualization methods of limited data sets have been used for clinical decision support.[3,4] Data from EHRs and other health related data, however, are displayed primarily through techniques that have been used for many years, e.g. fishbone diagrams for lab values, or by using charts and graphs. There have been few successful attempts to visualize massive amounts of disparate health care data.

Effective visualization techniques of large health data sets will allow users to see patterns they would not otherwise see. With many sources of health related data containing many parameters, the ability to visually explore the collective data has the potential to reveal valuable information.[5] There are many data elements and attributes in healthcare data. We propose that grouping and aggregating related data elements via a priori categorization (e.g. laboratory results or vital sign data) or data-driven methods (e.g. correlation) will facilitate developing visualization techniques that will allow users to see patterns in large data and elicit further inquiry of the data. We also believe the user should be able to further explore the data by opening the visual representation of a set of data elements to see trends representing aggregated data and drilling down even further to the subsets of the data. By having an interactive visualization, the ability to explore and gain a deeper understanding[6] of what the data represent will encourage adoption of the visualization technique, assuming the visual presentation minimizes cognitive burden.

Related Work

There are numerous reports in the literature related to data visualization in health care, most focusing on the technical aspects of visualization, medical imaging, and genomics. A number of prototypes have been also been reported. LifeLines, first described in 1996 by Plaisant and colleagues,[7,8] was used to visualize health data across a personal health record using timelines. Lifelines evolved to become Lifelines2, a visualization tool using categorical point event data across multiple records. More recently, Eventflow, similar to Lifelines2, also addresses the need to have a system to support interval events.[9]

Novel visualization techniques using EHRs was somewhat limited until 2009 when the HITECH Act mandated EHR implementation. In addition to evolving changes to LifeLines, several prototypes are in various stages of development. Most reported techniques are interactive, allowing the user to explore data incorporated as one visual display. For example, Zhang, et. al.[10] use a radial starburst visualization of multiple data points from one health record permitting users to drill down on data to single time points. Klimov and Shahar describe a prototype called VISITORS (Visualization of Time-Oriented Records) using time-oriented data sets with an interface to explore longitudinal values.[11] These approaches are similar to that we are taking, but we believe the historical queries and identification of the data elements and clusters will enhance visualization of relevant data.

Methods

Using historical data queries of Duke‘s EHR system (called DEDUCE) we will identify what data elements are in queries and classify them according to the types of information sought (e.g. outcomes, outliers of treatment methods). Groups of related data elements will be incorporated into a visualization that allows a quick comparison of the data from a large population with the ability to view trends over time within a chosen measure.

The following example illustrates our approach using the Department of Defense mock EHR data. We will look at aggregated health related data from an Army unit pre- deployment using visualization to discover differences within the group. We will then compare the same data elements post-deployment to identify changes. These time periods can be compared with the group later diagnosed with post-traumatic stress disorder to identify outliers and what data elements might have caused the outlier. In this example, Army personnel between 25-30 years old who have been deployed can be compared to the population of all adults in the system, or all Army or Marine or Navy personnel. We will employ visualization methods that show aggregated groups of data elements with a distribution per population, with the ability to drill down in the data and display longitudinal data for selected data elements.

The key to selecting the most effective method of visualization is to understand how to address the informational value of the data. We expect classes of data elements with the greatest variation to stand out. We will statistically pre-process data as an enhancement to visualization, eliminating null associations and unimportant variables (statistically). In comparing groups, the visualization method should clearly show differences. Further examination of data should also permit the easy application of different filters and the ability to hone down on subsets of data.

Discussion

1. Radial Coordinates Visualization. We have developed an initial multivariate visualization tool in the statistical programming environment R, using the RGL package to enable real-time interactive visualizations. This radial-coordinates visualization prototype is inspired by parallel-coordinate[12,13] and star plot[14] multivariate visualization techniques.

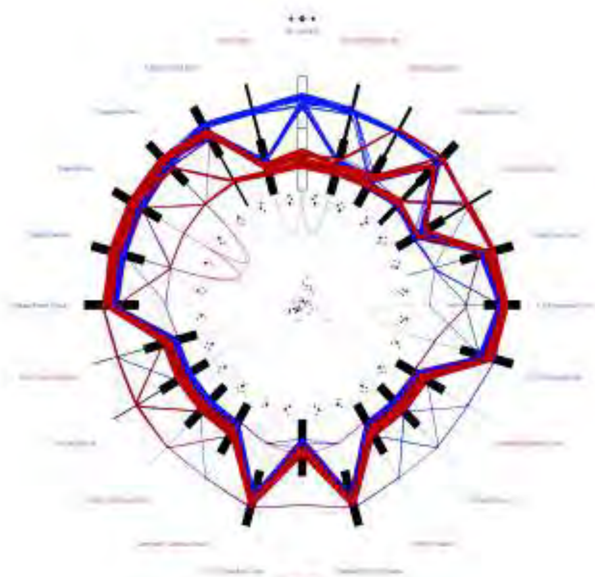


Figure 1: Radial coordinates visualization of DEDUCE queries.

Figure 1 shows an example radial coordinates visualization using queries from the top two users in Duke's DEDUCE EHR query tool. Each line represents a query, and the value for each axis represents how often that data element was used in the given query (typically zero or one). The lines are colored by system user. A circular layout of the axes has the advantage of a square aspect ratio when compared to standard parallel coordinates axes, which can be beneficial for large numbers of axes. Within this framework we have looked at additional improvements to standard parallel coordinates techniques, such as showing data distributions directly for each axis based on data type. For continuous data we display a box-and-whiskers plot (not shown in Figure 1), for discrete integer-valued data we display a histogram with bin width proportional to number of entities with that value, and for categorical data we display a stacked bar chart, with bar length proportional to number of entities with that value. This enables rapid evaluation of the various data types for a heterogeneous dataset, and of the distribution for each variable. In addition, we introduce line spreading to mitigate the problems of multiple lines collapsing to a single data point for discrete and categorical data, extending the parallel sets method (http://eagereyes.org/parallel-sets) to enable visualization of individual data entities, and the incorporation of non-categorical data.

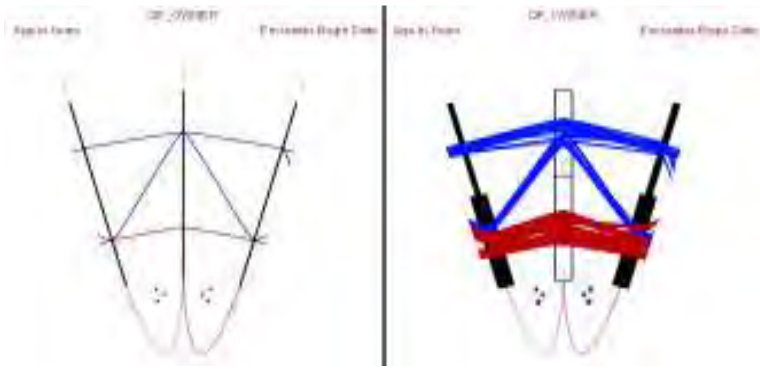


Figure 2: Line spreading (right) for discrete and categorical data enables improved visualization of multiple entities with the same value.

The close-up in Figure 2 illustrates the improvement possible using axis visualization with line spreading (right). The “QF_OWNER” categorical data (user 1 vs. user 2) is displayed using a stacked bar chart, with segment length proportional to number of entities with that value (there are slightly more queries from user 1), and individual lines are spread out within each bar segment based on their position on neighboring axes. The other two discrete integer-valued data elements are displayed using a histogram with bin width proportional to number of data entities with that value, and individual lines are spread out within each bin. With the visualization on the right it is much easier to follow individual lines between axes, and to see clustering of lines.

Axis-ordering is a well-known problem with parallel-coordinates techniques. We have experimented with a number of techniques for clustering axes based on correlation between axes. We also utilize correlation to flip axes to try to minimize line crossings, based on positive or negative correlation with a given axis. To enhance these techniques we also draw curved arcs connecting axis pairs, with opacity and line width proportional to correlation magnitude, and color based on correlation polarity (blue = negative, red = positive). Colored axis labels indicate whether the axis has been flipped (blue) or not (red).

The central space in the radial coordinates visualization enables the display of supplemental visualizations. Inspired by Holten and vanWijk,[15] we draw pair-wise scatterplots just below neighboring axes, and in the center we draw a scatterplot of the first two principal components. In the future, we plan to enable a number of different visualizations to be placed here, chosen interactively by the user. Each scatterplot and the radial coordinates visualization are linked together, such that selection in any visualization is reflected in the other visualizations.

2. Force-Directed Network Visualization. We have also developed a force-directed node-and-link network visualization to investigate queries from DEDUCE queries, implemented in the Processing programming environment. Figure 3 shows the same data as Figure 1, with individual query data elements drawn as circles, and de-identified system users (in this case the top two users of the system) as squares

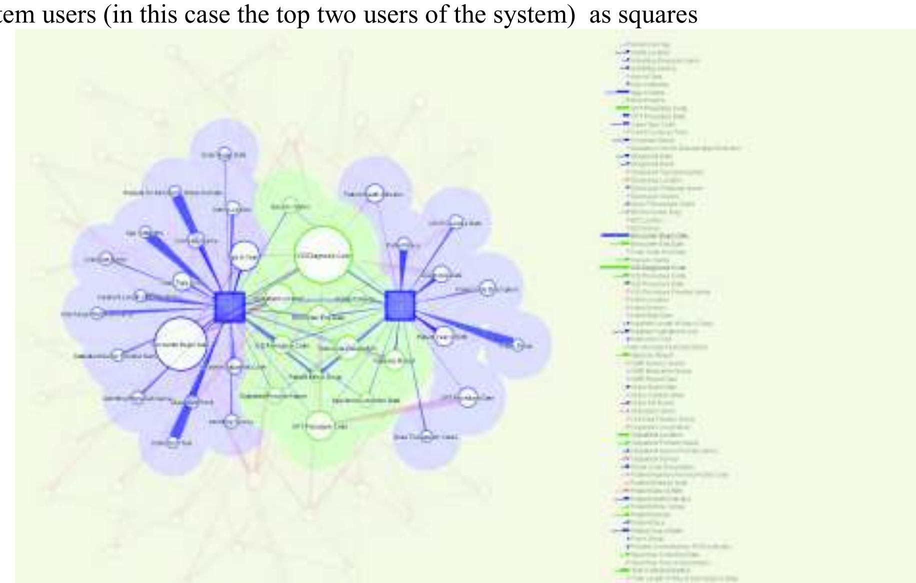


Figure 3: Force-directed layout visualization of DEDUCE queries.

The size of each circle represents how often it was used as a query element across all queries, and the size of each square represents the number of queries made by that user. Links between circles represent how often each element was used together in a series of queries, with each end scaled based on the relative importance at each end of the link. Links between circles and squares represent how often each user made a query on each element. Nodes are placed via a force-directed layout based on the overall strength of each link. In this example the user has highlighted the two users, which in turn highlights nodes connected to those users, while deemphasizing all other nodes. Nodes that are connected to both users are highlighted in green, whereas nodes that are connected to just one user are highlighted in blue. A full list of data elements is shown to the right, with horizontal lines representing the number of times each element was used across all queries (equivalent to circle size), and the number of other elements connected to. The user can interactively select nodes via the node-link diagram or the list of elements.

Some relationships are more easily discernible in one representation vs. the other. E.g. it is perhaps more readily apparent in Figure 3 that ICD Diagnosis Code is the most-used query element, and both users used that element, whereas in Figure 2 it is more apparent that Patient Gender, Patient Race, and Patient Diagnosis Date are all strongly correlated (i.e. they tended to be used together in queries), and that one of the users (red) included those elements more than the other. Our approach going forward will therefore combine such visualizations to enable multiple linked views of the data.

Conclusions

Compressing petabytes of health care data representing many data elements into various groups of related data presented visually with an interface that allows the user to interactively explore the data elements, to our knowledge never been done. There is the potential to detect causal relationships between various sets of data, which may lead to improved health care costs.

Acknowledgements
This work is supported by research funds from the Department of Defense, Award number W81XWH-13-1-0061.

References

1. Miranda ML, Edwards SE. Use of spatial analysis to support environmental health research and practice. NC Med J 2011;72:132-5.
2. Miranda ML, Edwards SE, Anthopolos R, Dolinsky DH, Kemper AR. The Built Environment and Childhood Obesity in Durham, North Carolina. Clin Pediatr (Phila) 2012.
3. Mane KK, Bizon C, Owen P, Gersing K, Mostafa J, Schmitt C. Patient Electronic Health Data–Driven Approach to Clinical Decision Support. Clinical and Translational Science 2011;4:369-71.
4. Mane KK, Bizon, C, Owne, P, Mostafa, J, Gersing, K and Schmitt, C. A Paradigm Shift: Electronic Health Records Data in Clinical Practice (Abstract). In: 2011 CTSA Annual Informatics Meeting. Natcher Conference Center, NIH Campus, Bethesda, MD; 2011:64-5.
5. Gershon N, Eick SG. Visualization's new tack: Making sense of information. Spectrum, IEEE 1995;32:38-40, 2, 4-7, 55-6.
6. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artif Intell Med 2006;38:115-35.
7. Plaisant C Milash B, Rose A, Widoff S, Shneiderman B. LifeLines: Visualizing Personal Histories. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1996:221-227.
8. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: using visualization to enhance navigation and analysis of patient records. *Proc. AMIA Symp.* 1998:76-80.
9. Lifelines2: Discovering Temporal Categorical Patterns Across Multiple Records. http://www.cs.umd.edu/hcil/lifelines2/. Accessed September 5, 2013.
10. Zhang Z, Wang B, Ahmed F, et al. The Five W's for Information Visualization with Application to Healthcare Informatics. *IEEE transactions on visualization and computer graphics.* Jun 3 2013.
11. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif. Intell. Med.* May 2010;49(1):11-31.
12. d'Ocagne M. Coordonnees parallels et axiale. Gautier-Villars, Paris 1885.
13. Inselberg A. The plane with parallel coordinates. *The Visual Computer*. 1(2):69-91.
14. Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth, 1983.
15. Holten and vanWijk, Evaluation of Cluster Identification Performance for Different PCP Variants, Computer Graphics Forum, 29(3), 793-802, 2010.

Appendix B

Demonstration of Visualization of EHR and Health Related Data for Information Discovery

David Borland[1] Vivian West[2] W. Ed Hammond[2]

[1] Renaissance Computing Institute, The University of North Carolina at Chapel Hill

[2]Duke Center for Health Informatics, Duke University

Introduction

In this demonstration we present research we are conducting in response to a program announcement solicited by the Assistant Secretary of Defense for Health Affairs, Defense Health Program. We have developed visualization prototypes for multivariate heterogeneous data along with visualizations of retrospective data queries from DEDUCE, an electronic health record (EHR) query tool developed at Duke University.

Our current approach involves incorporating data queries of Duke‘s EHR system to help identify what data elements are used in queries and classify them according to what types of information users were seeking (e.g. queries searching for outcomes, or outliers of treatment methods). Eventually groups of related data elements will be incorporated into a visualization that allows a quick comparison of the data from a large population with the ability to view trends over time within a chosen measure.

Methods

We have developed two interactive visualization prototypes, one a radial coordinates visualization (Figure 1, left) based on parallel coordinates techniques, and one a force-directed node-and-link network visualization (Figure 1, right). Our radial coordinates visualization is a multivariate visualization suitable for heterogeneous data that incorporates multiple supplemental scatterplots, direct visualization of axis correlations, and a novel technique for spreading lines to enable improved visualization of individual lines and line clusters. Our force-directed network visualization enables the interactive selection of nodes to see relationships between groups of nodes.

In our demonstration we will show how various relationships in the data are reinforced between the two views, and how different visualizations can be more adept at showing different relationships in the data. We will also apply these visualization techniques to publicly available EHR data from the NHS.

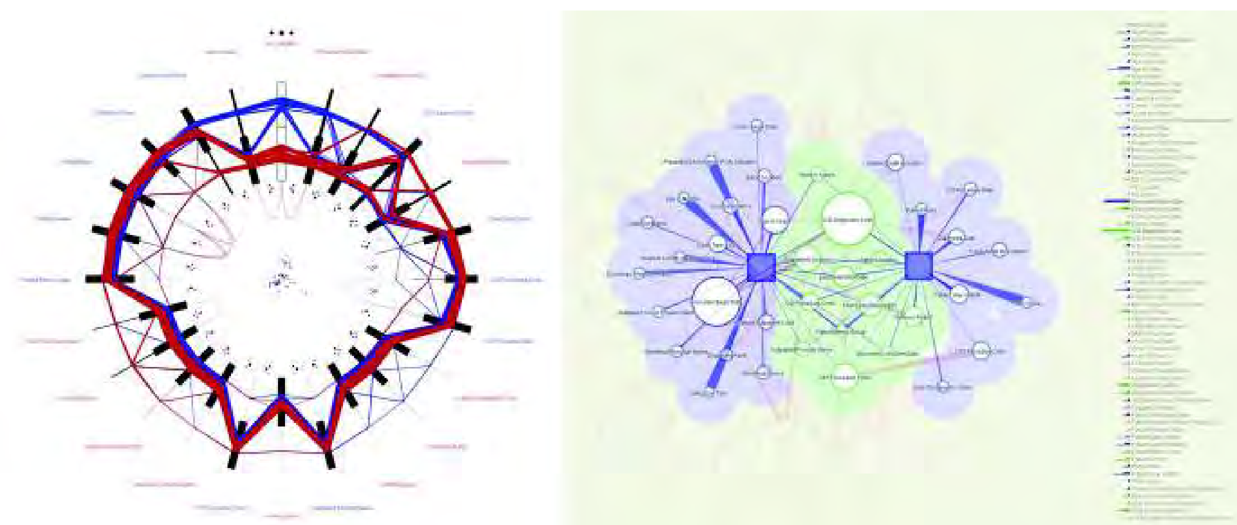
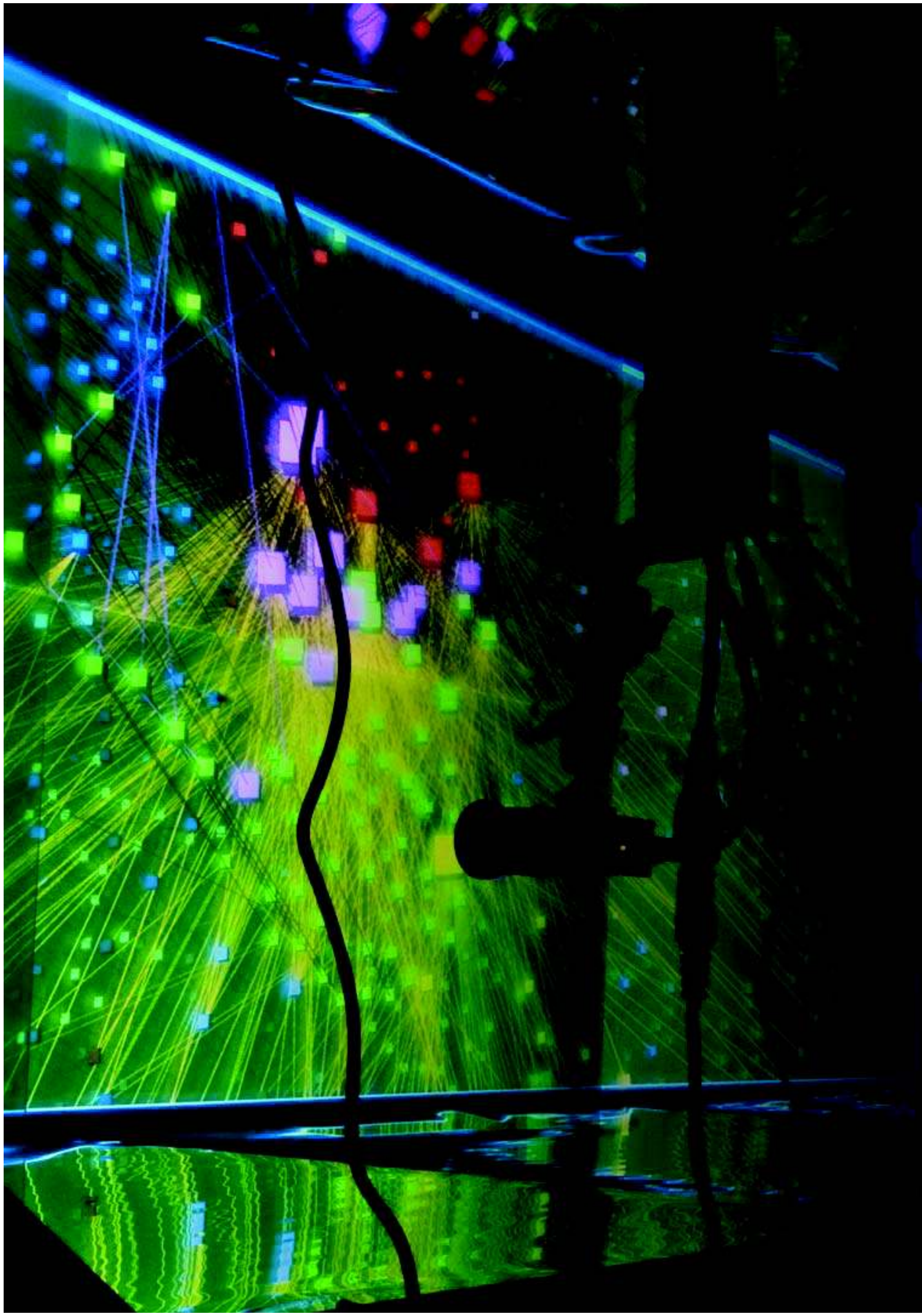


Figure 1: Radial coordinates (left) and force-directed network (right) visualizations of the same EHR query data

Acknowledgements

This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation

Appendix C



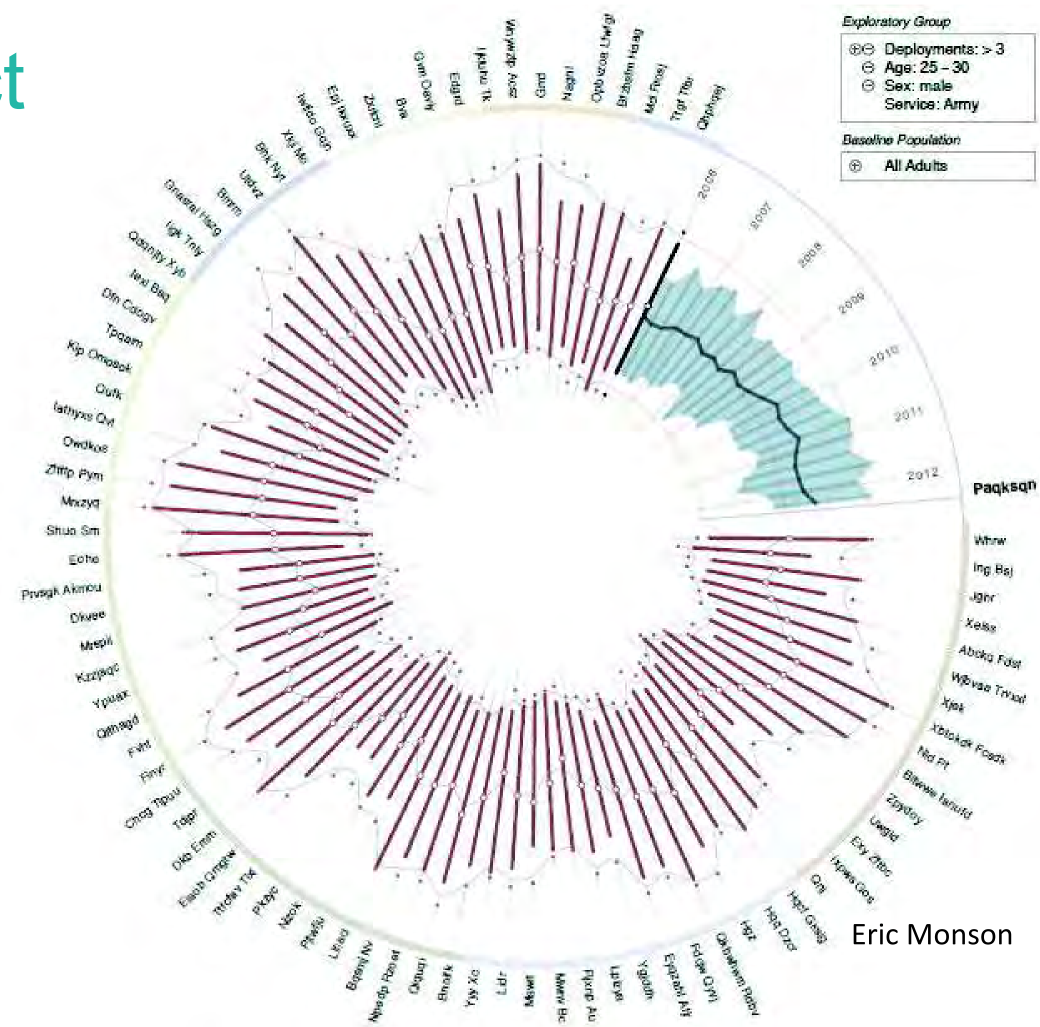
Visualization of Health Informatics Data

David Borland
Senior Visualization Researcher

RESEARCH ENGAGEMENT INNOVATION

Novel Visualization of Large Health Related Data Sets

- DoD-funded project
  - February 2013-August 2014
- Duke Center for Health Informatics (DHCI)
  - Ed Hammond (PI)
  - Vivian West



Eric Monson

The Problem

- Electronic Health Data
  - Amount and types rapidly expanding.
- Visual Exploration
  - Reveal patterns in the data
  - Elicit further inquiry
  - Lead to valuable new knowledge

The Data

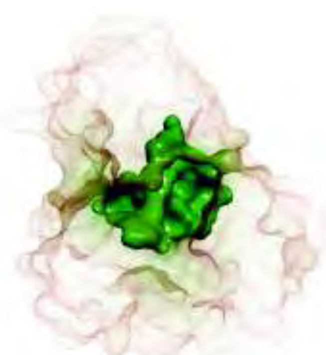
- DEDUCE Query Data
  - Duke Enterprise Data Unified Content Explorer
- NHS Data
  - Practice and primary care trust (PCT) data from the UK's NHS
- DOD Synthetic EHR Data
  - 10,000 patient synthetic data set

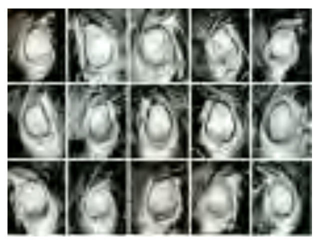
The Data

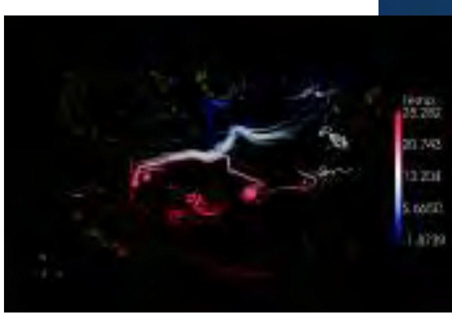
- DEDUCE Query Data
  - Duke Enterprise Data Unified Content Explorer
- NHS Data
  - Practice and primary care trust (PCT) data from the UK's NHS
- DOD Synthetic EHR Data
  - 10,000 patient synthetic data set

Scientific vs. Information Visualization

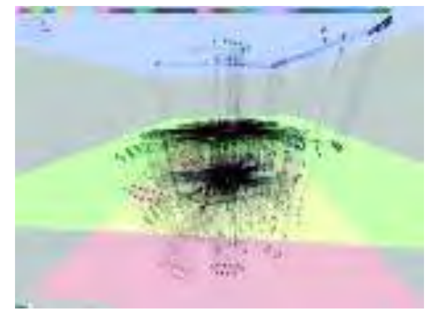
- Scientific Visualization
  - Spatially-embedded data
- Information Visualization
  - Abstract data





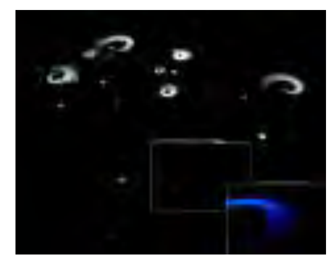


Me



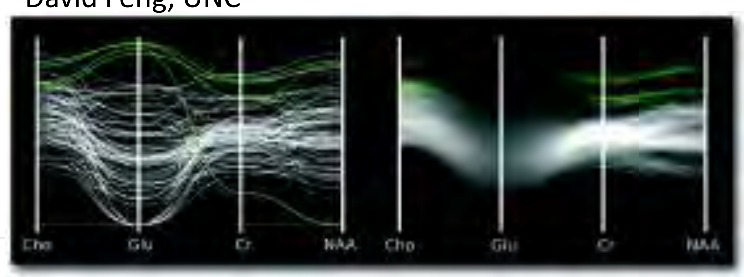
Hong Yi, RENCI

Jeff Heard, RENCI



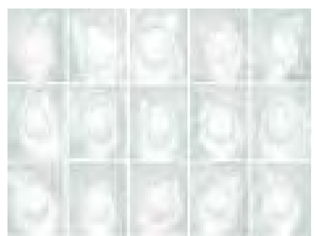
Jeff Heard, RENCI

David Feng, UNC

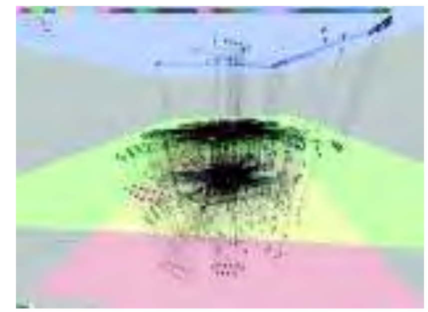


Scientific vs. Information Visualization

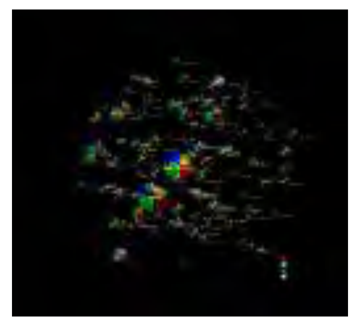
- Scientific Visualization
  - Spatially-embedded data



- Information Visualization
  - Abstract data



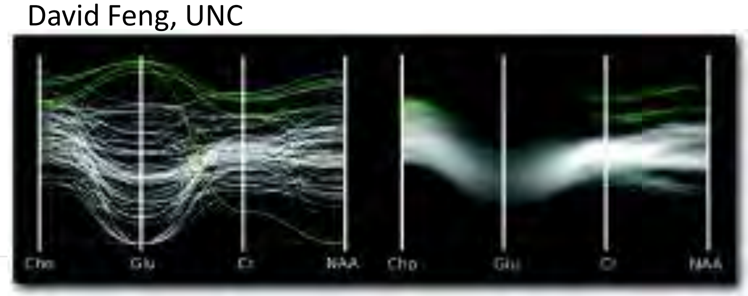
Hong Yi, RENCI



Jeff Heard, RENCI

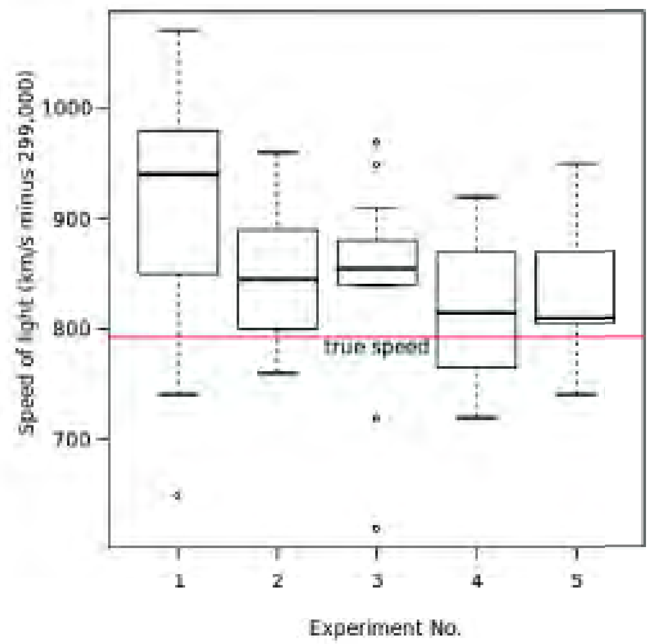
Jeff Heard, RENCI

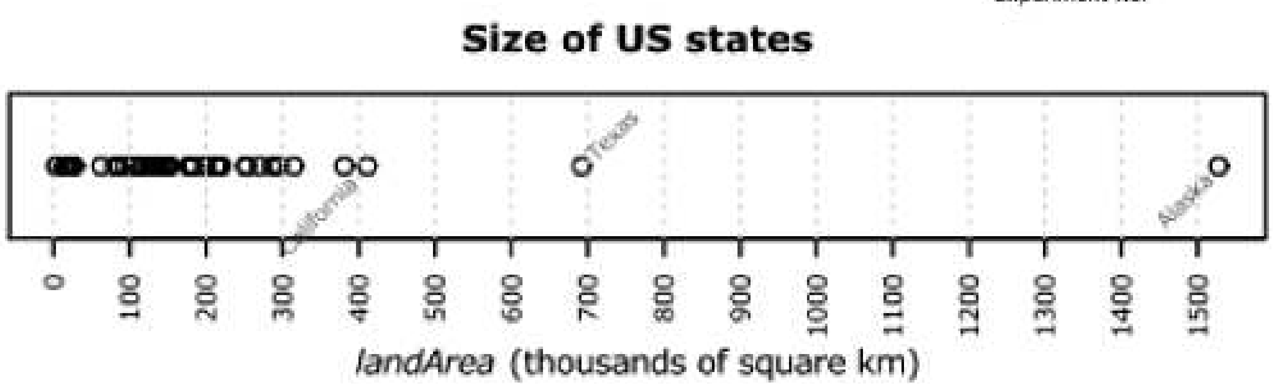
David Feng, UNC



Univariate Visualization

- Single variable
  - Bar chart
  - Line graph
  - Box plot
  - Etc.

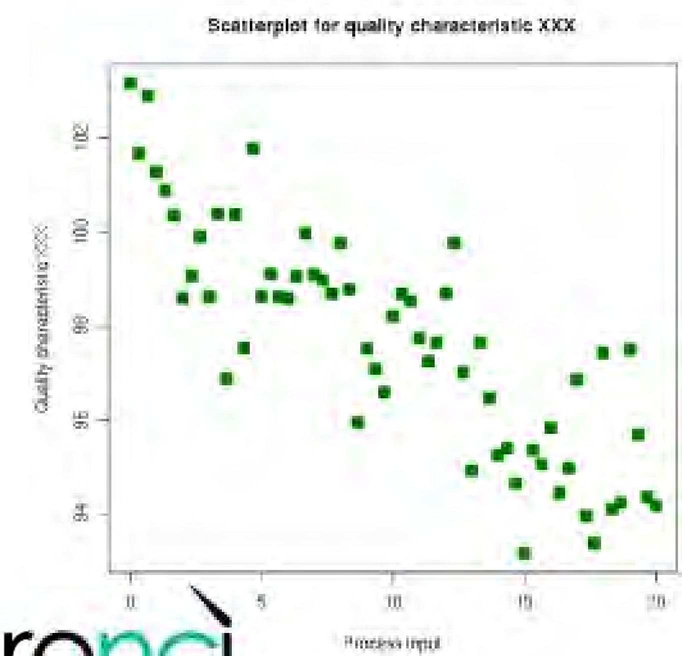


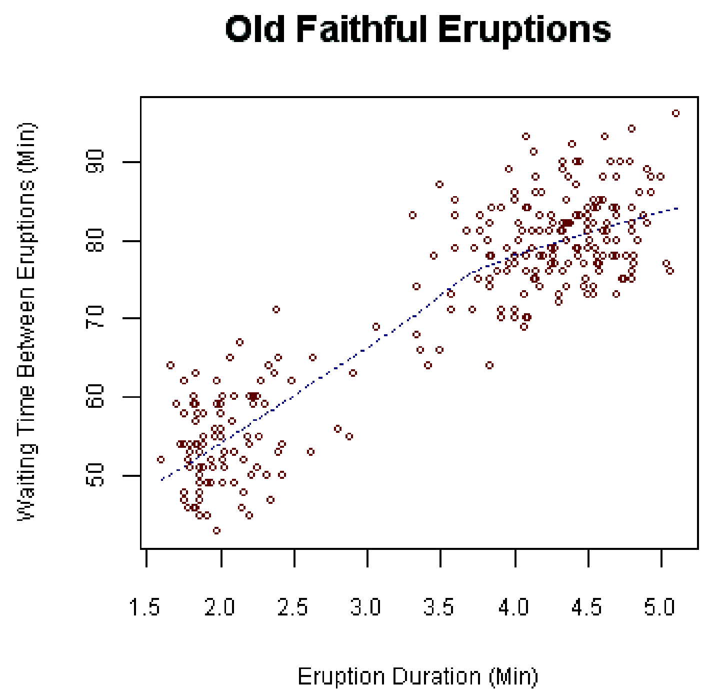


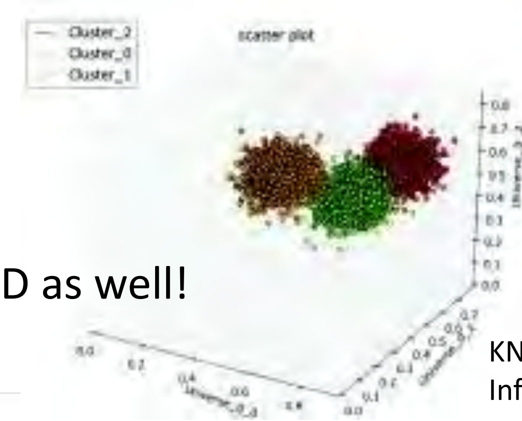
renci

Bivariate Visualization

- Scatter plot
  - Show clusters
  - Show correlation







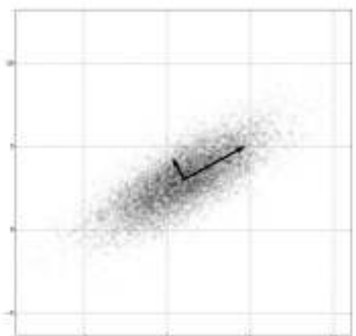
3D as well!

KNIME Konstanz Information Miner

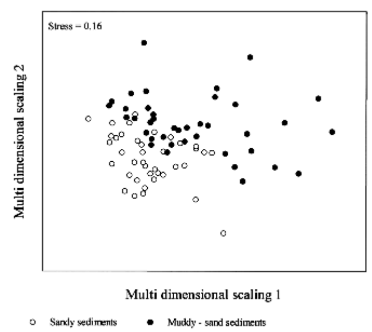
Multivariate Visualization: Two Basic Strategies

Dimension Reduction

Principle components analysis (PCA)



Multidimensional scaling (MDS)



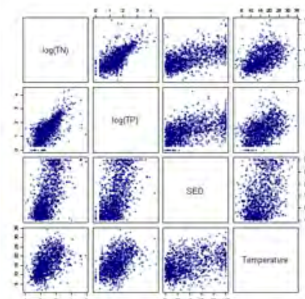
Jaramillo et al.

Self-organizing maps (SOM)

Sarat Kocherlakota

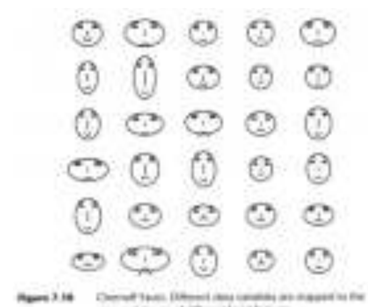
Direct Visualization

Scatter plot matrices



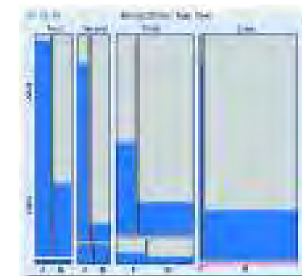
EPA

Glyphs



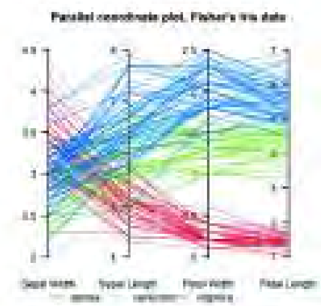
Colin Ware

Mosaic plots



Martin Theus

Parallel coordinates



Star plots

Also statistical techniques: clustering, etc.

Dimension Reduction

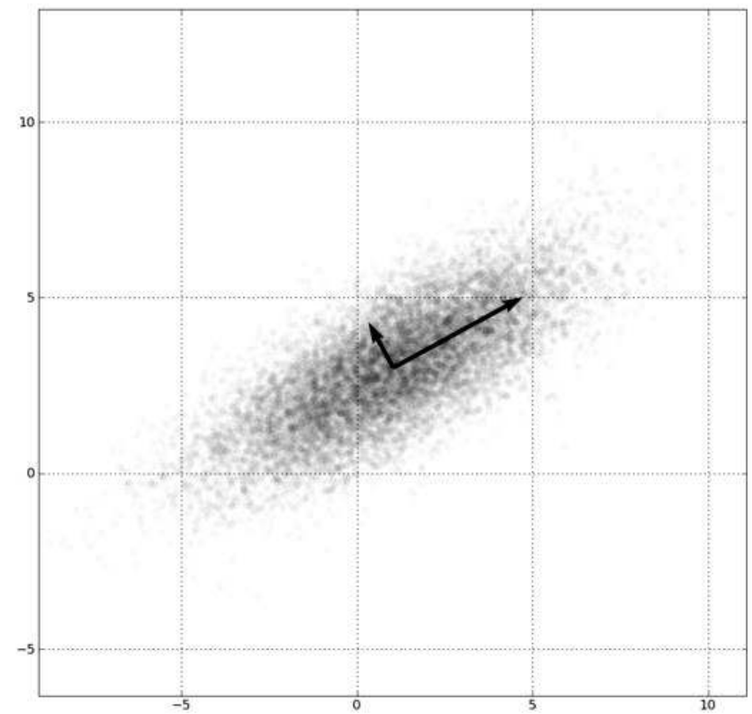
- Find lower dimensional representation that maintains some of the higher-dimensional structure
  - Typically 2D or 3D
  - Also grand tour…

- Use standard 2D or 3D visualization techniques
  - E.g. scatter plots

renci

Principle Components Analysis (PCA)

- Karl Pearson, 1901
- Set of orthogonal axes
  - Maximize variance
- Project into lower-dimensional representation
  - Hopefully contains much of the variation in the data
  - Always losing something…



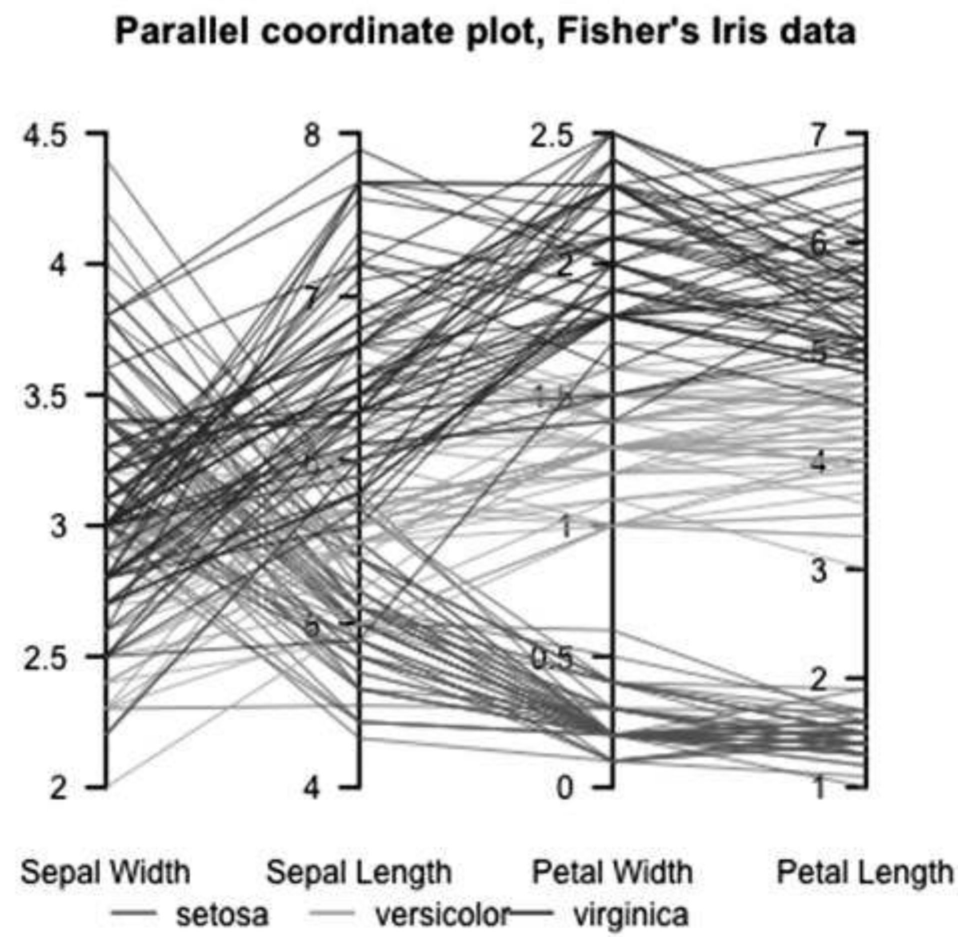
renci

Direct Visualization

- Try to directly represent all (or a high-dimensional subset) the data dimensions
  - Mapping multiple dimensions to visually salient features
  - Reorganizing dimensions

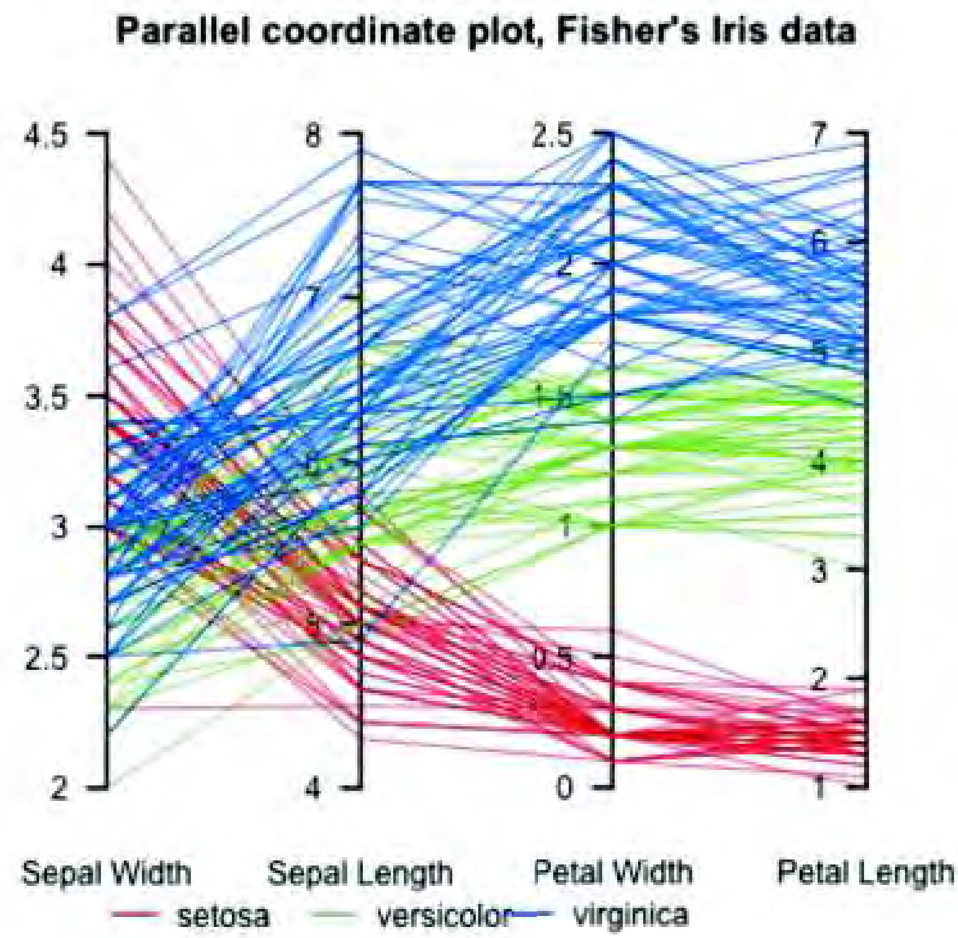
Parallel Coordinates

- Henry Gannett, 1880
- Each dimension a parallel axis
  - Each data point a line
  - Axis ordering important
- Interaction important
  - Brushing



Parallel Coordinates

- Henry Gannett, 1880
- Each dimension a parallel axis
  - Each data point a line
  - Axis ordering important
- Interaction important
  - Brushing



EHR Data

- Many entities
  - Individuals, practices, PCTs
- Many variables/dimensions
  - Continuous
  - Discrete
  - Categorical

PROBLEM:

IRB Approval

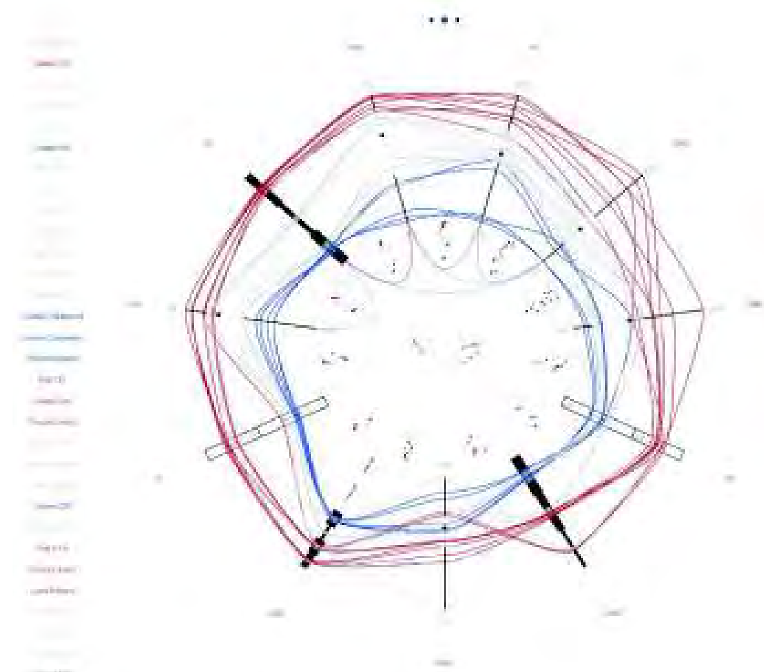
SOLUTION:

Non-health-
related data sets

R Matey

V. West, D. Borland, and W. E. Hammond. **Visualization of EHR and health related data for information discovery**. *AMIA Workshop on Visual Analytics in Healthcare*. 2013.

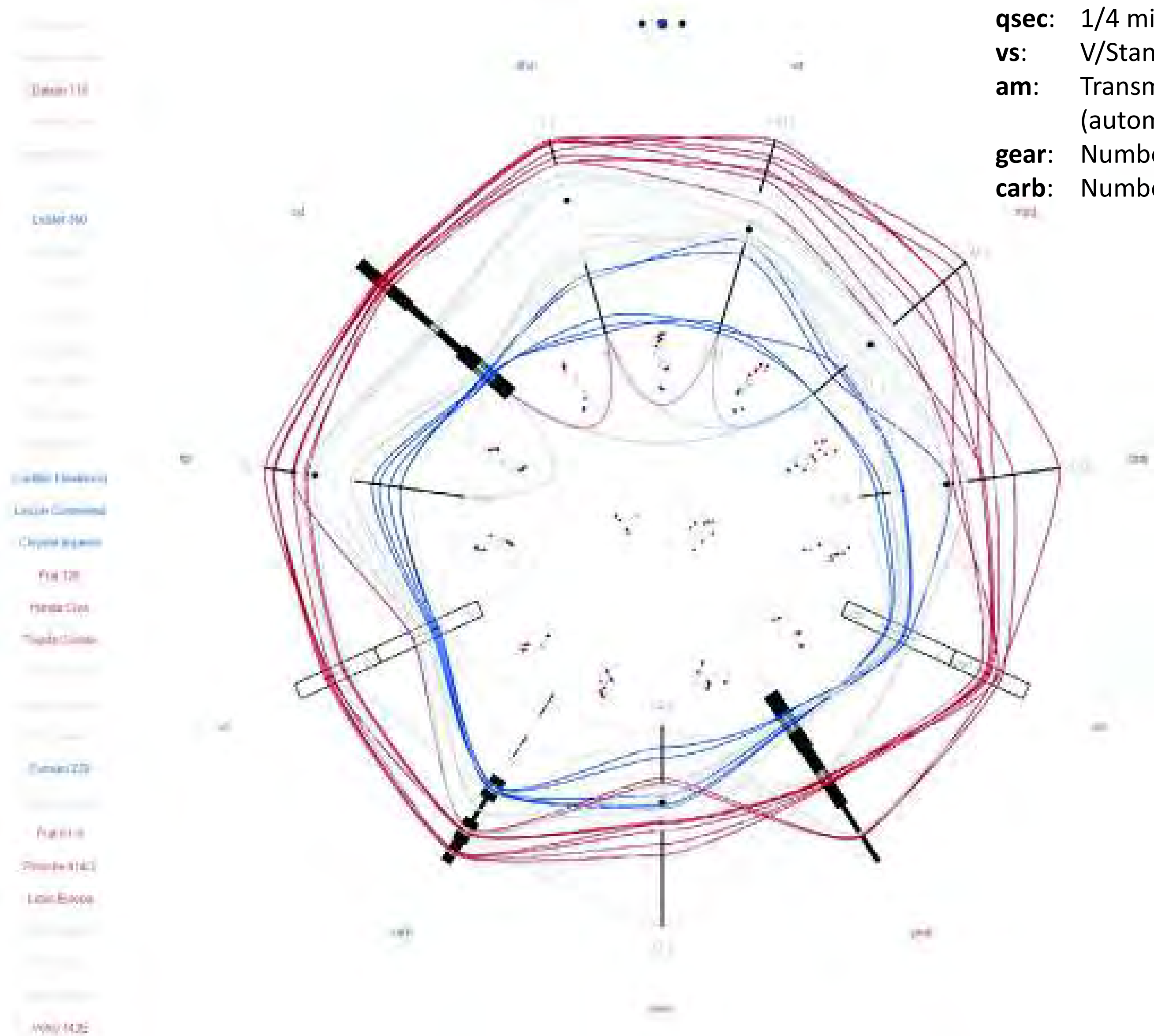
- Initial prototype developed in R
  - Access to statistical methods
    - prcomp: principle components analysis
    - cor: correlation matrix
  - Interactive graphics
    - Rgl
  - Built-in data sets
    - mtcars
      - 32 entities
      - 11 variables



renci

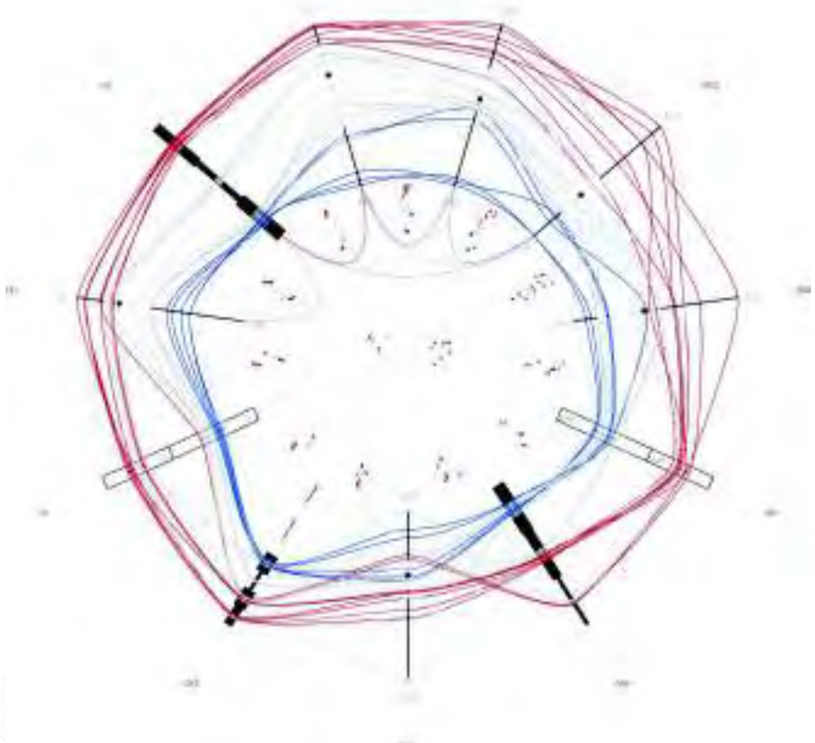
Radial Coordinates Prototype

mpg: Miles/(US) gallon
cyl: Number of cylinders
disp: Displacement (cu.in.)
hp: Gross horsepower
drat: Rear axle ratio
wt: Weight (lb/1000)
qsec: 1/4 mile time
vs: V/Standard
am: Transmission (automatic, manual)
gear: Number of forward gears
carb: Number of carburetors

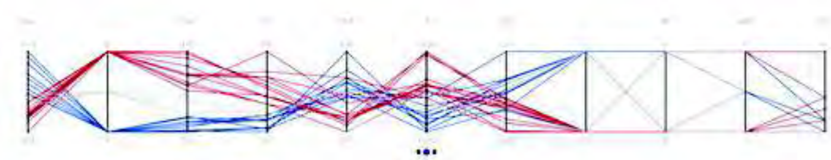


Features

- Radial axis layout
  - Maintains square aspect ratio, even with many dimensions

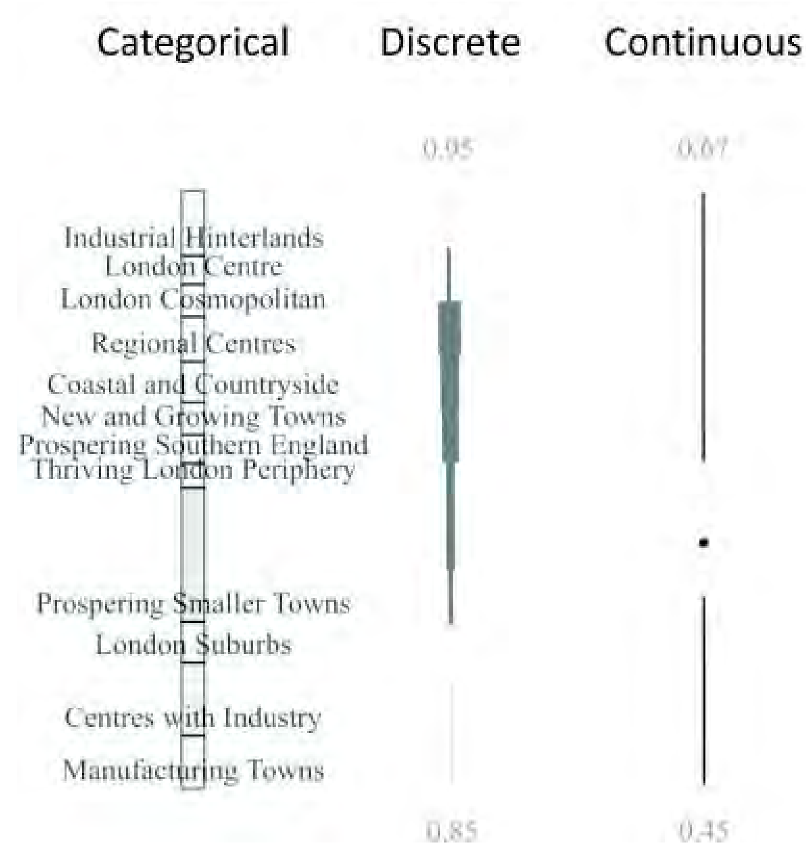


vs



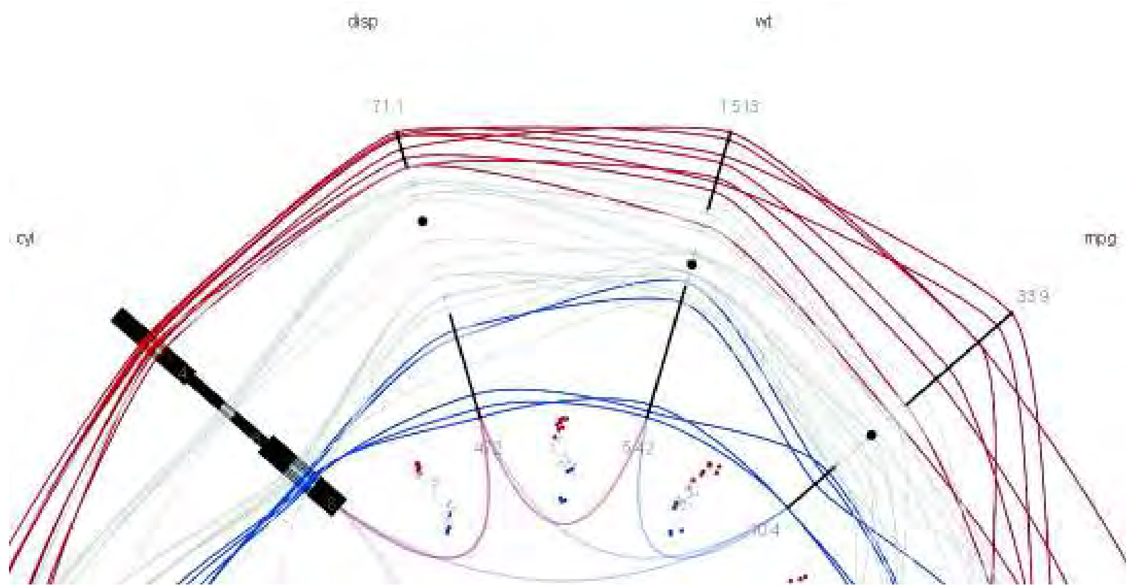
Features

- Data-type dependent axis distribution visualization
  – Categorical: Stacked bar chart
  – Discrete: Histogram
  – Continuous: Tufte-style box and whiskers plot
    • Quartiles and median



Features

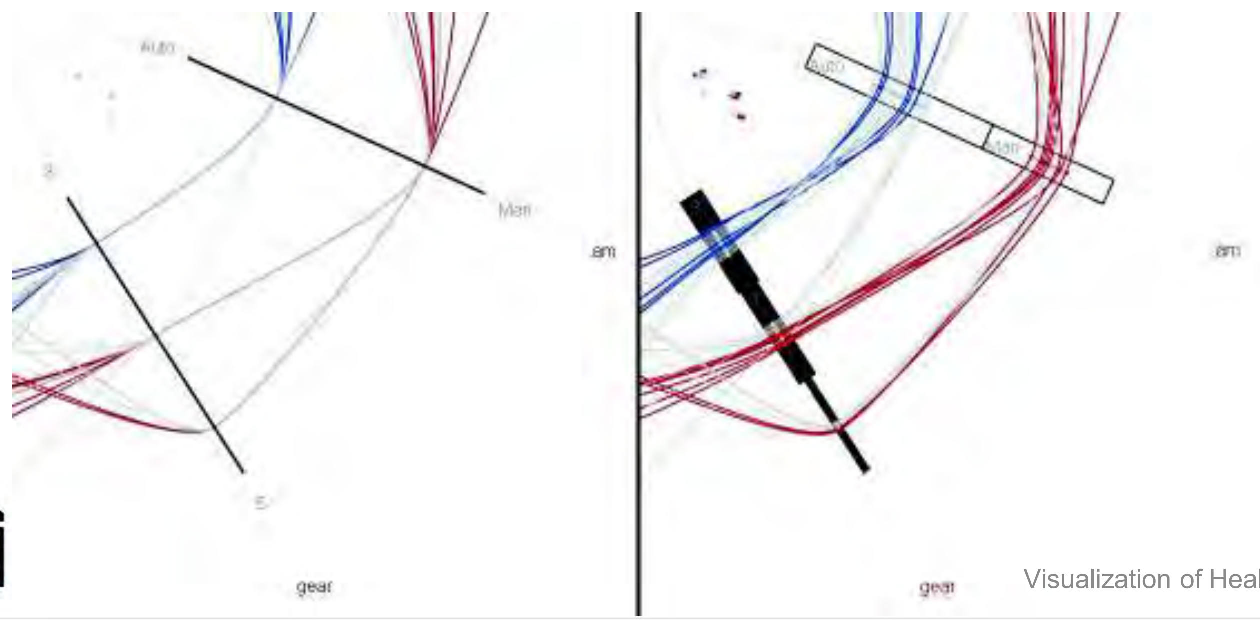
- Axis clustering
  - Based on correlations
- Axis flipping
  - Based on correlations
- Correlation visualization
  - Arcs connecting axes, (red positive, blue negative)



Features

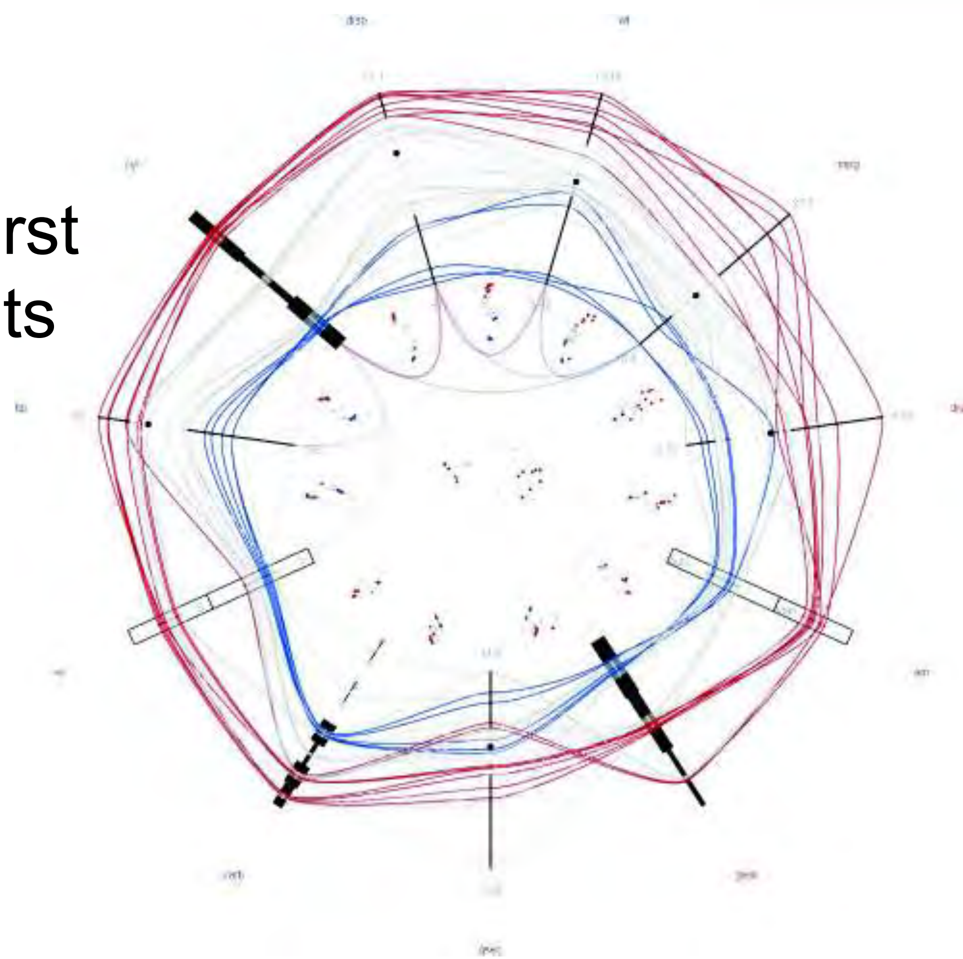
Graham and Kennedy, **Using Curves to Enhance Parallel Coordinates Visualisations**, *Proc. Information Visualization*. 2003.

- Curved lines
  - Easier to follow, cross axes at right angle
- Line spreading
  - Categorical and discrete axes



Features

- Scatterplots
  - Neighboring axes and first two principle components
- Linked brushing
  - Curves
  - Points



The Good, the Bad, and the Ugly

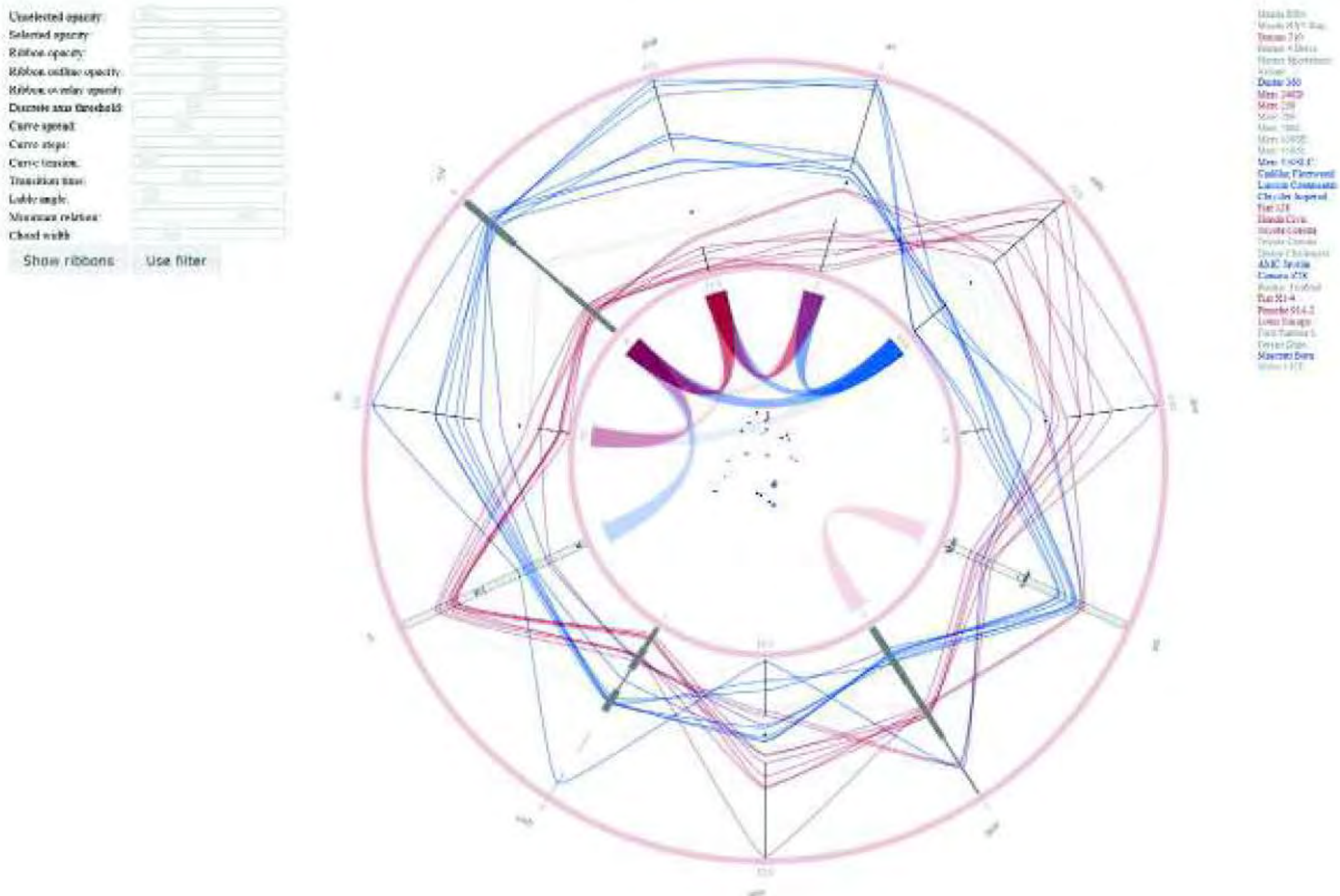
- The Good
  - Enables identification of clusters and relationships between variables in multidimensional datasets
  - R has a lot of useful statistical functions
  - rgl enables some degree of interactivity
- The Bad
  - Initial processing/rendering very slow
  - Advanced selection difficult
  - No clear path for a GUI
- The Ugly
  - Blended transparency doesn't work as expected

d3 to the rescue?

- Data-driven documents (d3)
  - Javascript library for web-based visualization
  - Map data to webpage's document object model (DOM)
    - Typically scalable vector graphics (SVG)
  - Apply event-driven transformations
    - Built for interactivity

renci

Radial Coordinates in d3

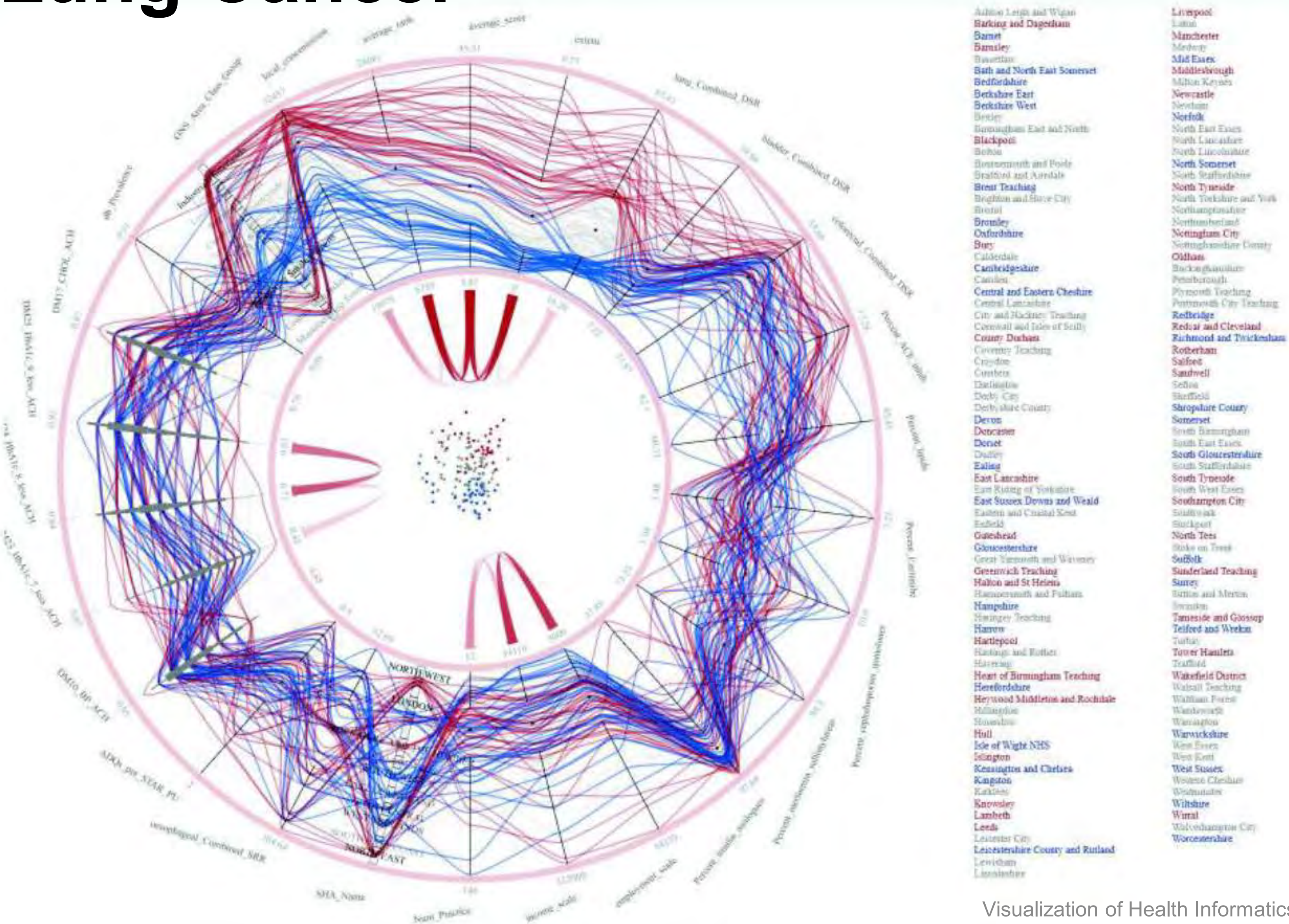


NHS Data

- Primary Care Trust
  - Regional collections of practices
  - 147 of the 152 PCTs in England
  - 26 variables/dimensions

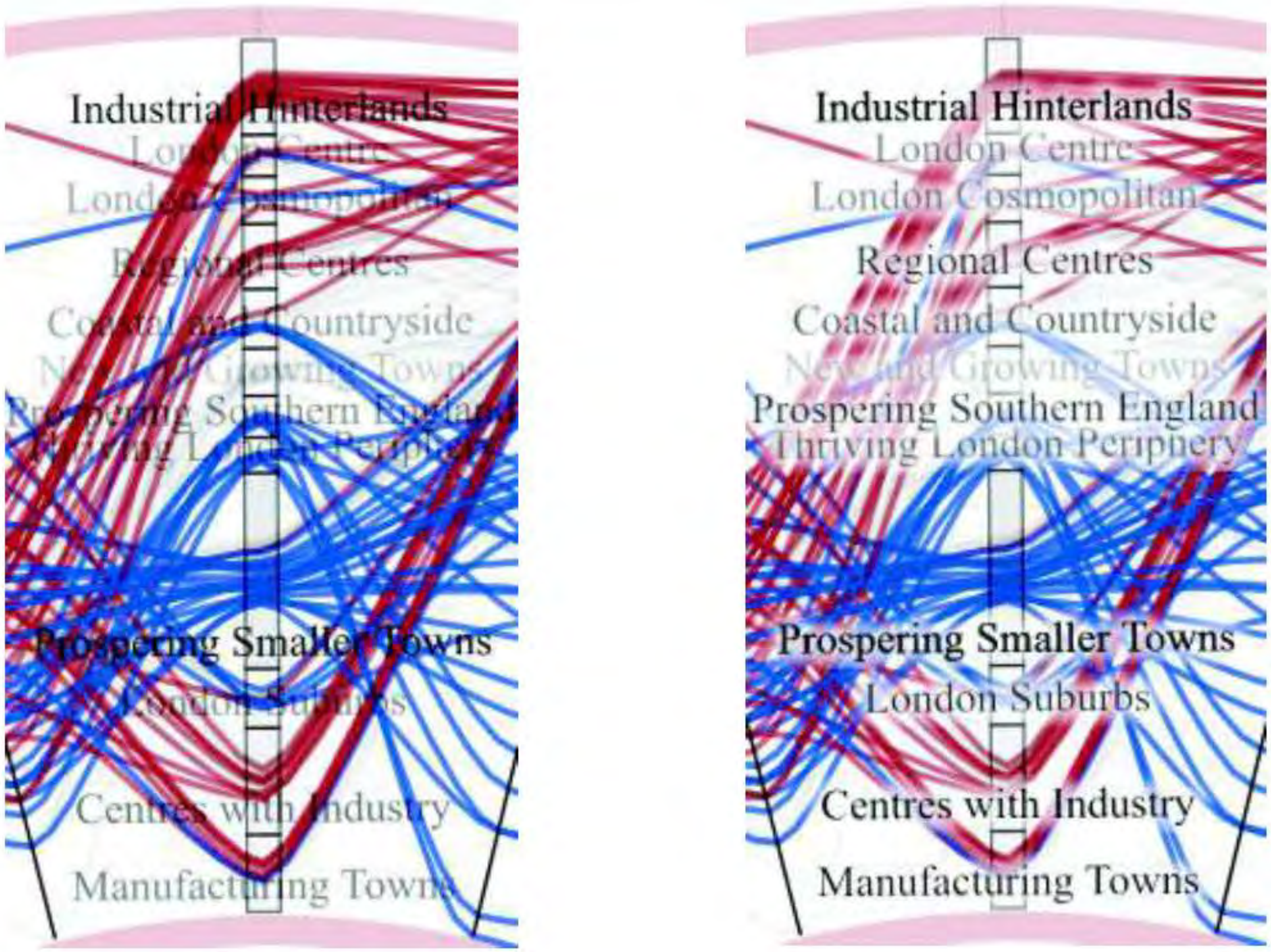
renci

Lung Cancer



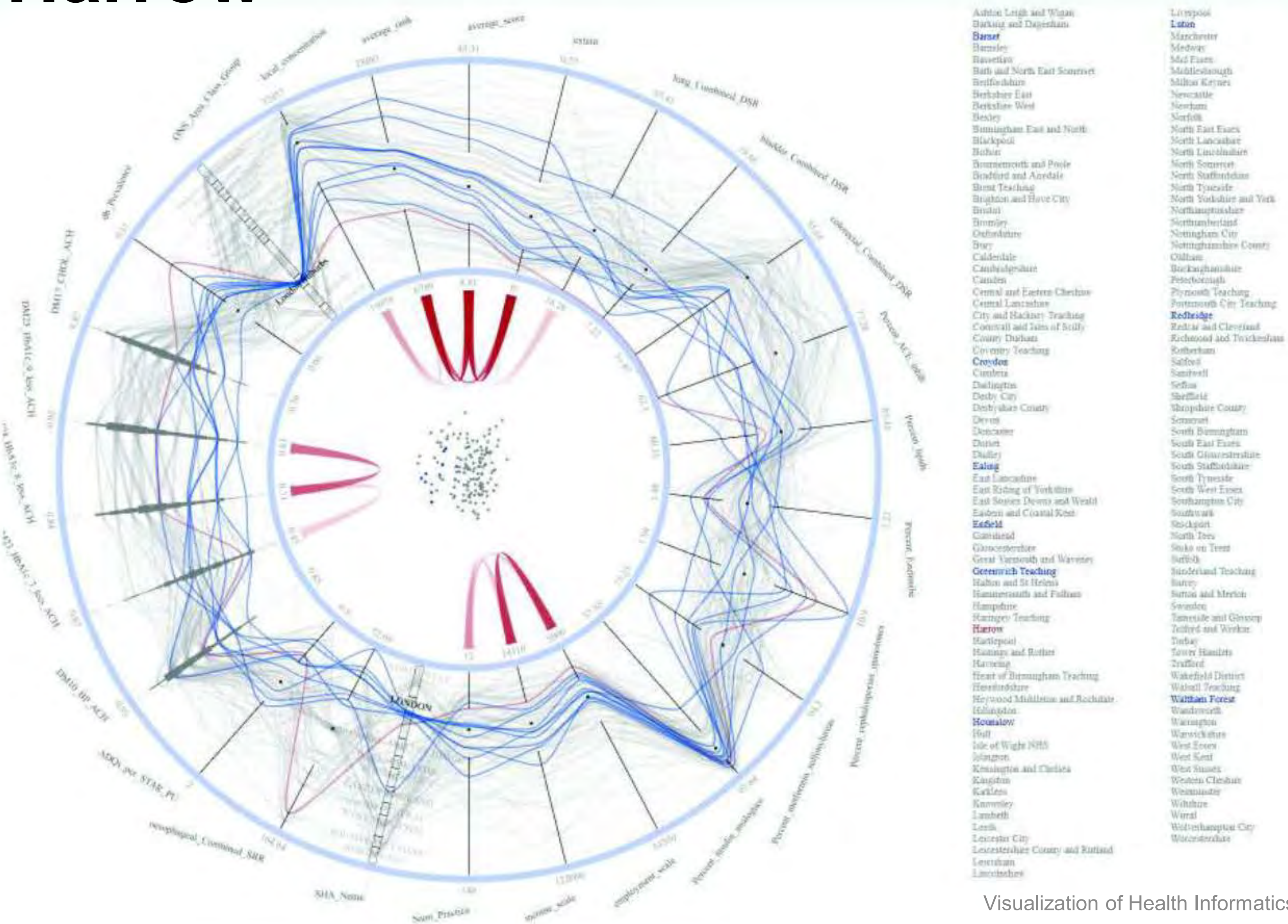
NORTH WEST
LONDON
NORTH EAST
SHA_Name
Lung_Combined_DSR
Bladder_Combined_DSR
Percent_lipids
Percent_Cardarelli
Percent_ACE_inhib
income_scale
employment_scale
Liverpool
Manchester
Mid Essex
Middlesbrough
Newcastle
Norfolk
North Somerset
North Tyneside
Nottingham City
Oldham
Redbridge
Redcar and Cleveland
Richmond and Twickenham
Rotherham
Salford
Sandwell
Shropshire County
Somerset
South Gloucestershire
South Tyneside
Southampton City
North Tees
Suffolk
Sunderland Teaching
Surrey
Tameside and Glossop
Telford and Wrekin
Tower Hamlets
Wakefield District
Warwickshire
West Sussex
Wiltshire
Wirral
Worcestershire
Barking and Dagenham
Barnet
Barnsley
Bath and North East Somerset
Bedfordshire
Berkshire East
Berkshire West
Blackpool
Brent Teaching
Bromley
Oxfordshire
Bury
Cambridgeshire
Central and Eastern Cheshire
County Durham
Devon
Doncaster
Dorset
Ealing
East Lancashire
East Sussex Downs and Weald
Gateshead
Gloucestershire
Greenwich Teaching
Halton and St Helens
Hampshire
Harrow
Hartlepool
Heart of Birmingham Teaching
Herefordshire
Heywood Middleton and Rochdale
Hull
Isle of Wight NHS
Islington
Kensington and Chelsea
Kingston
Knowsley
Lambeth
Leeds
Leicestershire County and Rutland

Lung Cancer by Classification



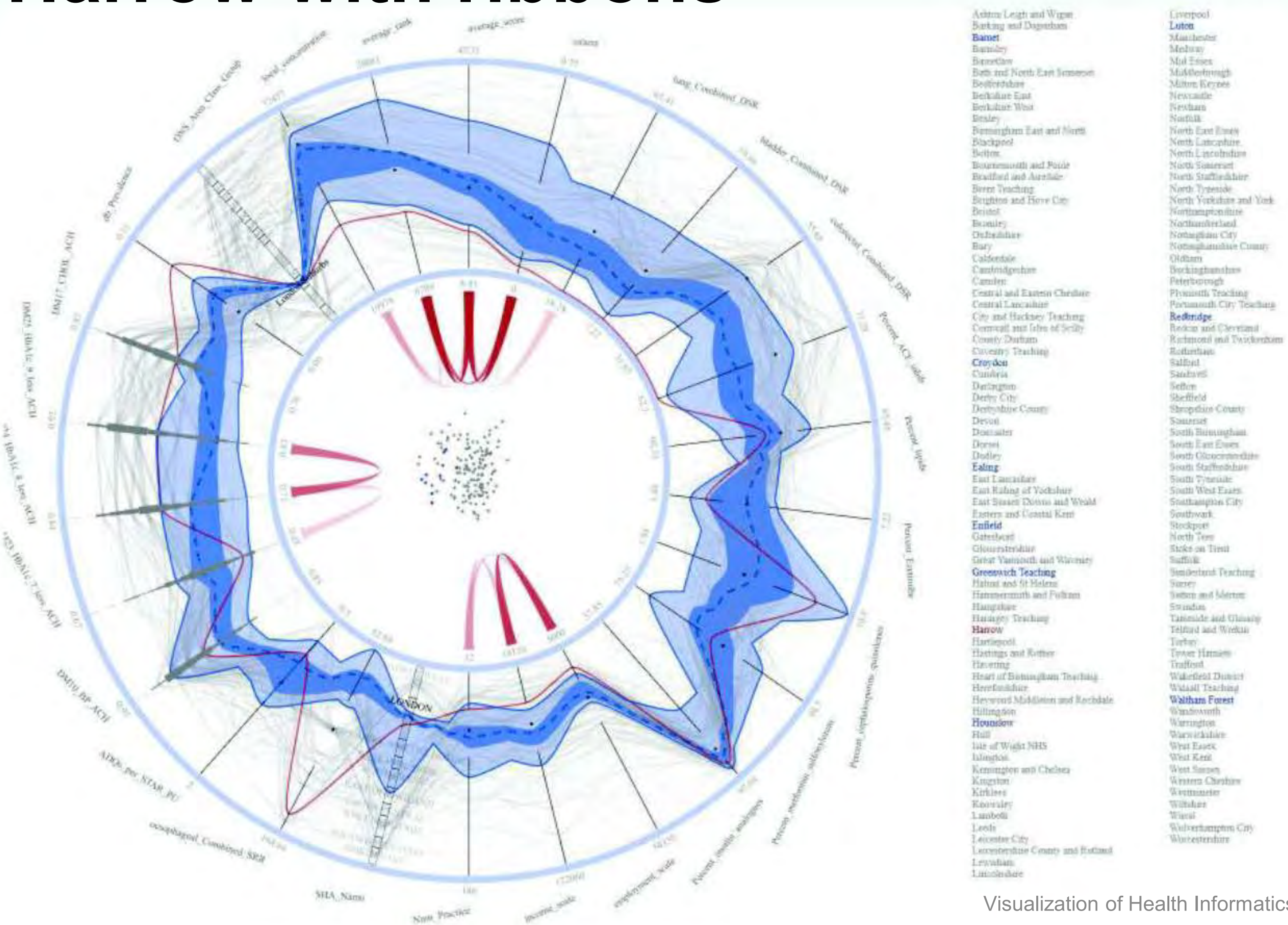
ONS_Area_Class_Group

Harrow

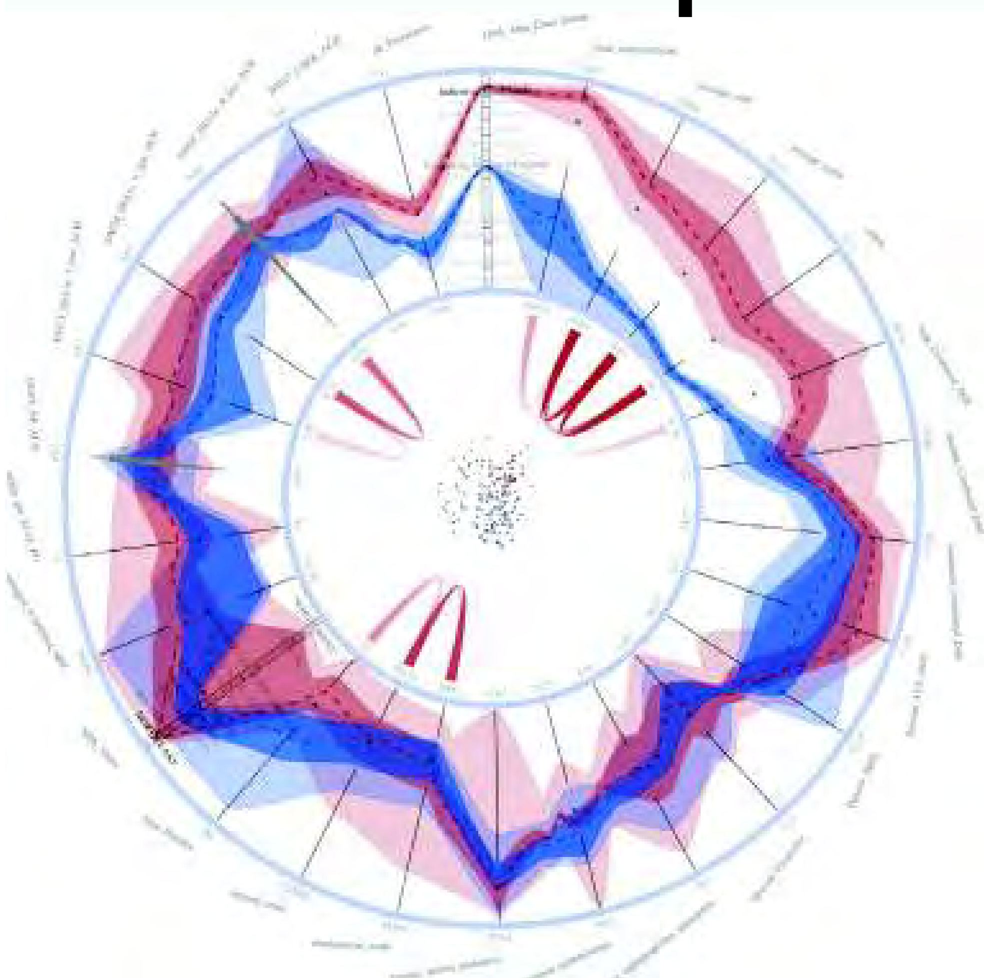


average_rank
average_score
lung_Combined_DSR
bladder_Combined_DSR
colorectal_Combined_DSR
Percent_ACE_inhib
Percent_lipids
Percent_Ezetimibe
Percent_cephalosporin_quinolones
Percent_metformin_sulfonylureas
Percent_insulin_analogues
employment_scale
income_scale
Num_Practice
SHA_Name
oesophageal_Combined_SRR
ADQs_per_STAR_PU
DM10_BP_ACH
DM25_HbA1c_9_less_ACH
DM17_CHOL_ACH
db_Prevalence
ONS_Area_Class_Group
local_concentration
LONDON
Ashton Leigh and Wigan
Barking and Dagenham
Barnet
Barnsley
Bassetlaw
Bath and North East Somerset
Bedfordshire
Berkshire East
Berkshire West
Bexley
Birmingham East and North
Blackpool
Bolton
Bournemouth and Poole
Bradford and Airedale
Brent Teaching
Brighton and Hove City
Bristol
Bromley
Buckinghamshire
Bury
Calderdale
Cambridgeshire
Camden
Central and Eastern Cheshire
Central Lancashire
City and Hackney Teaching
Cornwall and Isles of Scilly
County Durham
Coventry Teaching
Croydon
Cumbria
Darlington
Derby City
Derbyshire County
Devon
Doncaster
Dorset
Dudley
Ealing
East Lancashire
East Riding of Yorkshire
East Sussex Downs and Weald
Eastern and Coastal Kent
Enfield
Gateshead
Gloucestershire
Great Yarmouth and Waveney
Greenwich Teaching
Halton and St Helens
Hammersmith and Fulham
Hampshire
Haringey Teaching
Harrow
Hartlepool
Hastings and Rother
Havering
Heart of Birmingham Teaching
Herefordshire
Heywood Middleton and Rochdale
Hillingdon
Hounslow
Hull
Isle of Wight NHS
Islington
Kensington and Chelsea
Kingston
Kirklees
Knowsley
Lambeth
Leeds
Leicester City
Leicestershire County and Rutland
Lewisham
Lincolnshire
Liverpool
Luton
Manchester
Medway
Mid Essex
Middlesbrough
Milton Keynes
Newcastle
Newham
Norfolk
North East Essex
North Lancashire
North Lincolnshire
North Somerset
North Staffordshire
North Tyneside
North Yorkshire and York
Northamptonshire
Northumberland
Nottingham City
Nottinghamshire County
Oldham
Oxfordshire
Peterborough
Plymouth Teaching
Portsmouth City Teaching
Redbridge
Redcar and Cleveland
Richmond and Twickenham
Rotherham
Salford
Sandwell
Sefton
Sheffield
Shropshire County
Somerset
South Birmingham
South East Essex
South Gloucestershire
South Staffordshire
South Tyneside
South West Essex
Southampton City
Southwark
Stockport
North Tees
Stoke on Trent
Suffolk
Sunderland Teaching
Surrey
Sutton and Merton
Swindon
Tameside and Glossop
Telford and Wrekin
Torbay
Tower Hamlets
Trafford
Wakefield District
Walsall Teaching
Waltham Forest
Wandsworth
Warrington
Warwickshire
West Essex
West Kent
West Sussex
Western Cheshire
Westminster
Wiltshire
Wirral
Wolverhampton City
Worcestershire

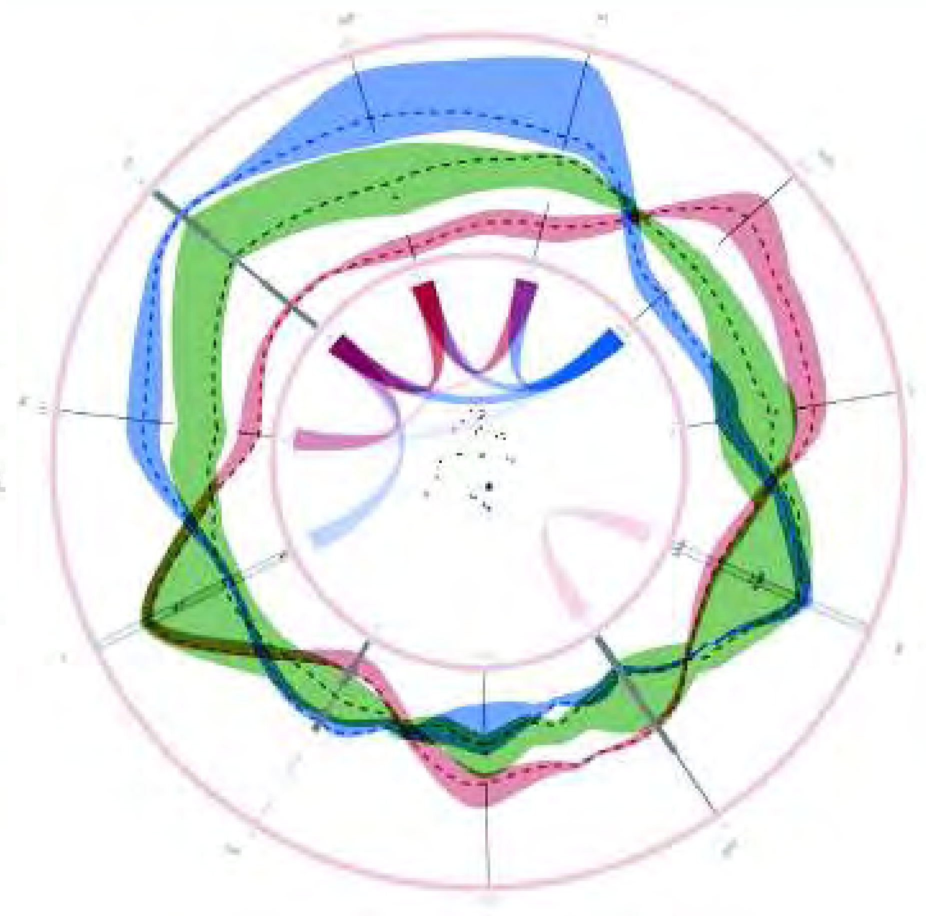
Harrow with ribbons



Ribbon examples



NHS PCT Data



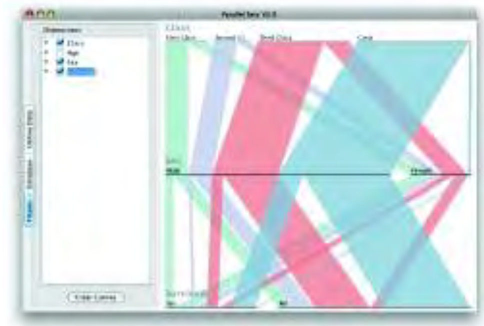
mtcars

Current Issues

- Big Data
  - Works with 10,000 patient data set, but very sluggish
  - Issues with blending
- No R
  - Don’t have access to useful statistical functions
  - Currently exporting multiple files per dataset from R
- Categorical ordering
  - Correlations currently based on arbitrary ordering
  - Doesn’t work well with ribbons

Future Work

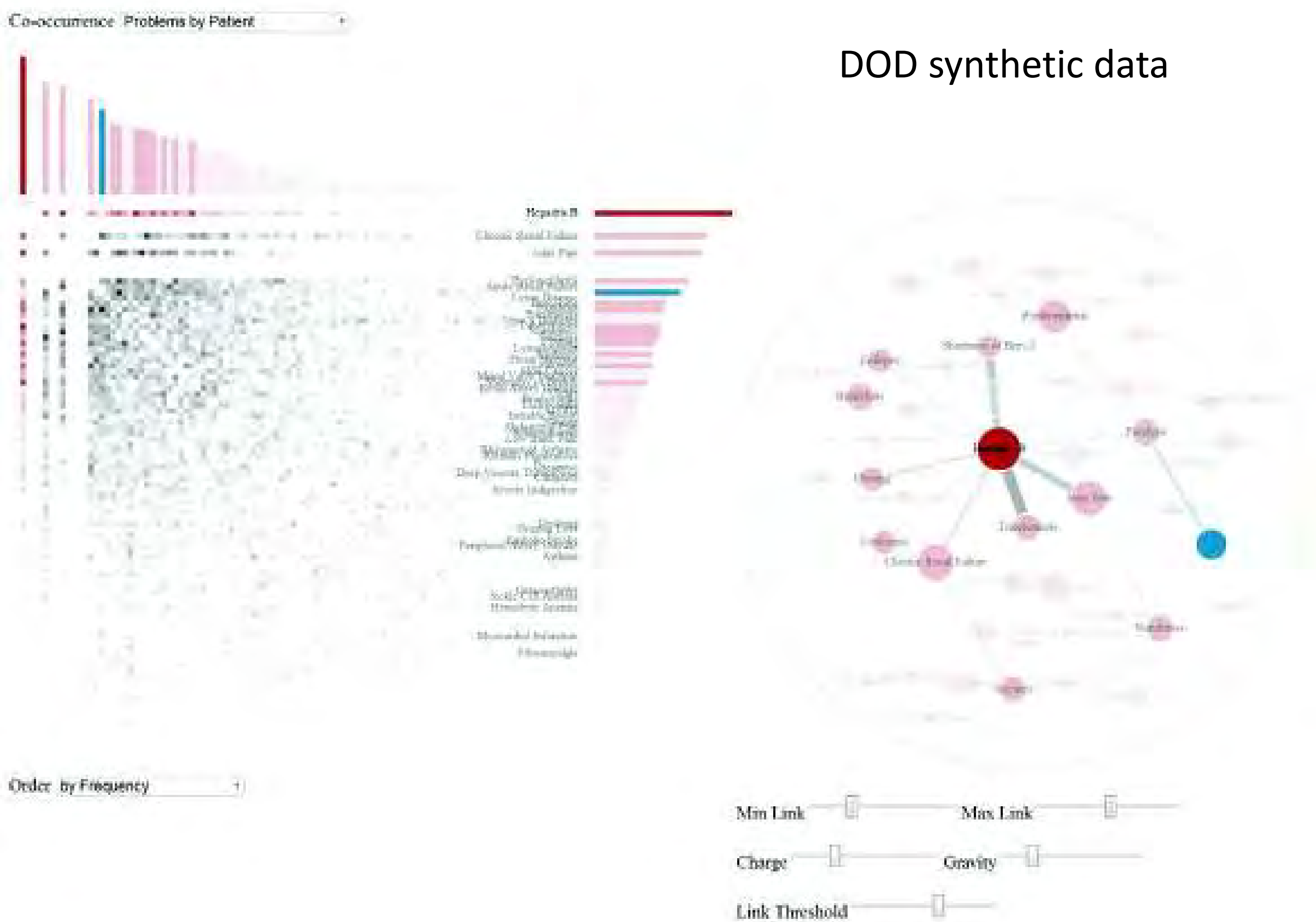
Bendix R, Kosara R, Hauser H. Parallel sets: **Visual analysis of categorical data**. *2005 IEEE Symposium on Information Visualization (InfoVis 2005)*. 2005.



- Incorporate parallel sets techniques for big data
  - Similar to ribbons
- Use Shiny
  - Library for interactive web applications using R
  - Combine with d3
- Dynamic categorical axis ordering

Future Work: Co-Occurrence Visualization

DOD synthetic data



Acknowledgments

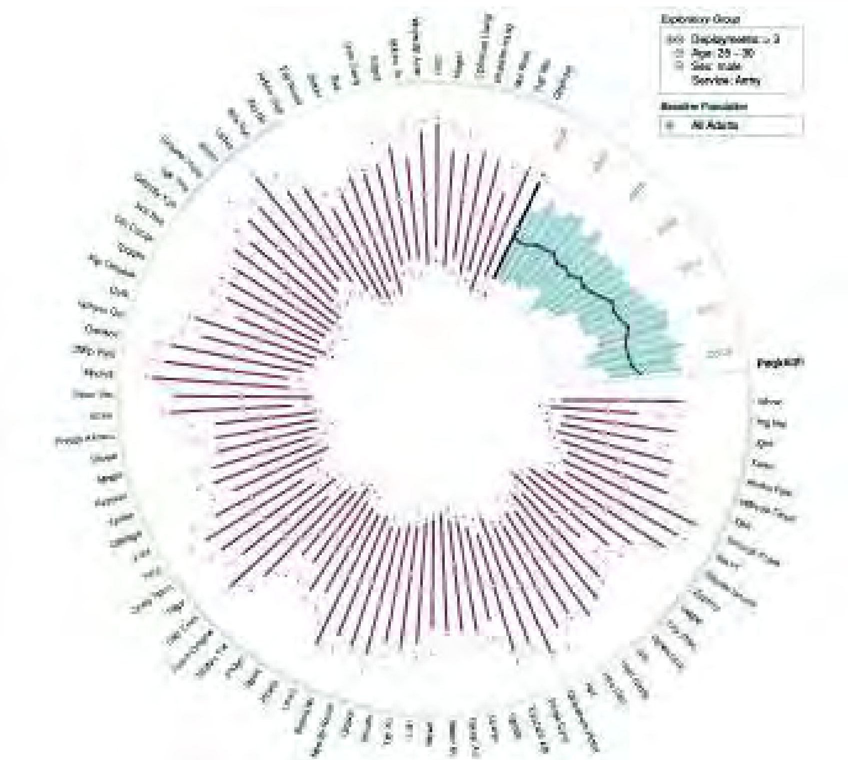
- Ed Hammond, Vivian West, and Rene Hart from DCHI
- Steve Evans, Genie Hinz, and Igor Akushevich
- Neil Stine: National Health Service and other UK data was made available courtesy of the BT Health Cloud.

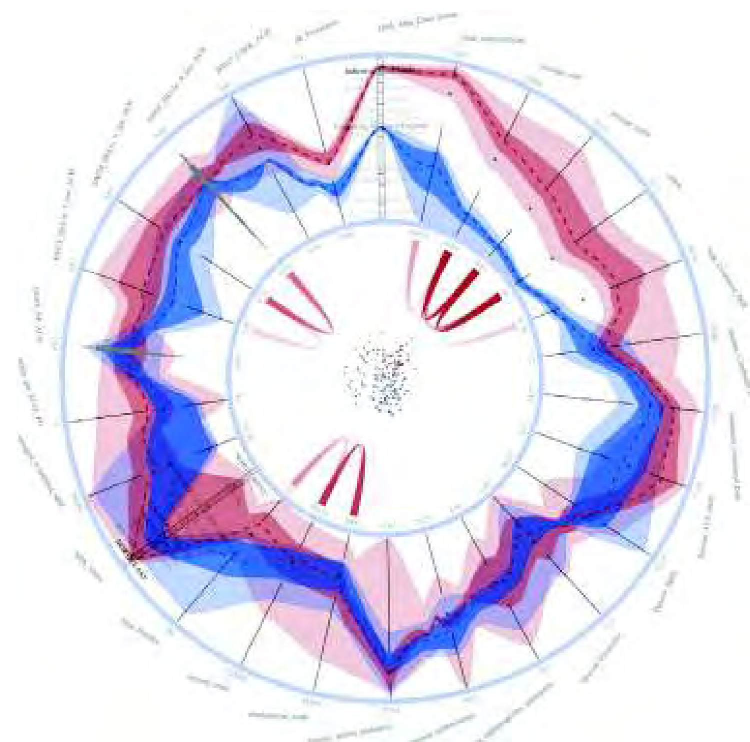
  – Work supported by the US Army Medical Research and Material Command (USARMC) under Grant No. W81XWH-13-0061. The views, opinions and/or findings contained in this presentation/report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

renci

Thanks!

- Any questions?





Appendix D

Demonstration of Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels

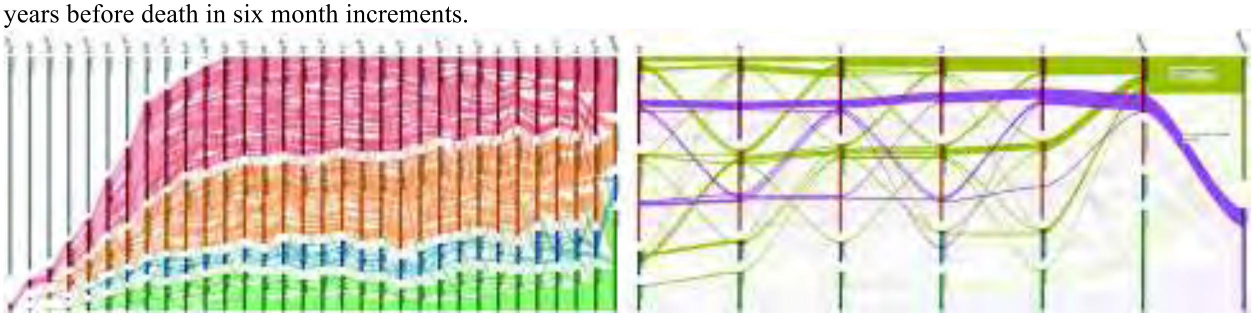
Hina Shah[1], David Borland[2], Eugenia McPeek Hinz[3], Vivian L. West[1], W. Ed Hammond[1]
[1]Duke Center for Health Informatics, Duke University, Durham, NC;
[2]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[3]DHTS Duke Medicine, Durham, NC

Introduction

In this demonstration we present a visualization tool for a cohort of patients with diabetes (via ICD9 codes) from Duke University's data warehouse, visualizing their Hemoglobin A1c (HbA1c) levels over time, aligned by death, to explore trajectories of glycemic control. To the best of our knowledge, temporal visualization of glycemic control for a diabetic population standardized on death has not previously been presented. Our visualization groups HbA1c values into ordered categories of glycemic control, utilizing a method based on parallel sets and Sankey diagrams to view temporal patterns in HbA1c values. We incorporate a number of features for interactive data exploration like: viewing the progression of values either forwards or backwards in time, highlighting multiple subpopulations, coloring based on the category along each path in the data or at the beginning/end of each path, and the incorporation of demographic data, such as gender.

Methods

Data from Duke University's data warehouse were extracted using DEDUCE, an electronic health record (EHR) query tool developed at Duke University. The final cohort includes data from 121 patients with diabetes mellitus (with and without complications), a death indicator, prescribed antihyperglycemics, and at least 10 years of HbA1c laboratory values. We average HbA1c values over 6 month time intervals. In the case of missing HbA1c values within a 6 month period, we first attempt to impute a HbA1c value from average glucose (AG) values over that period of time if available, otherwise the previous HbA1c value (measured or imputed) is carried forward. HbA1c values are then categorized based on the severity of diabetes: Normal < 5.7, Borderline [5.7, 6.5), Controlled [6.5, 8), and Uncontrolled ≥ 8. The sampled data is time-aligned by the death event for each patient. The visual representation of diabetes progression propagates backwards in time initially. Time is represented as number of years before death in six month increments.



(a) (b)

Figure 1. Diabetes progression visualizations without and with a gender axis: (a) Overview visualization with paths colored by the current HbA1c at each time step, useful for emphasizing overall temporal trends. (b) Adding a gender axis and selecting two groups, we can compare the variability of males who were Uncontrolled at death (olive) to women who were Uncontrolled at death (purple). Men appear to have more variability over the 5-year period being visualized, as shown by the large number of transitions between different categories.

Our visualization tool was developed using the D3 JavaScript library. The aim of this visualization is to investigate temporal trajectories of HbA1c levels for a large cohort of diabetes patients over a number of years prior to death. Parallel sets is chosen for showing HbA1c summary trajectories. Each vertical axis is a time step. The user can choose the frequency of these time steps, with a minimum sampling frequency of six months, and also the maximum number of years before death. The death event axis is placed at the right with all other time steps moving backwards in time to the left (Figure 1). Each vertical axis is split into the four HbA1c categories (Normal in green, Borderline in blue, Controlled in orange, and Uncontrolled in red), and a Missing category in grey (for patients with more than 10 years of data). The height of each axis category represents the proportion of the patients in that category at that point in time. Paths moving between axes recursively split, moving backwards from death to show the trajectories of similar groups of patients. The visualization can show trends either starting at the death event, i.e. going backwards in time, or starting at the last year in the visualization, i.e. going forward in time. The user can highlight one or more groups of patients by clicking on categories or trajectories to highlight the behavior of that group of patients going backward and forward in time, reducing visual clutter (Figure 2). We also include the ability to incorporate demographic data, such as gender, as additional axes (Figure 1). This feature enables the comparison of trajectories for different subpopulations based on data other than just HbA1c levels.

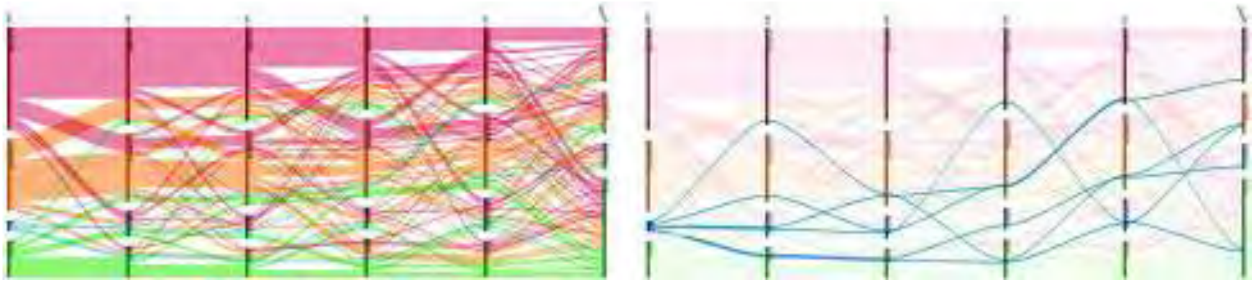


Figure 2. A 10-year range of data, sampled every two years, with forward propagation to show how the trajectories of patients change moving forward in time (left). Highlighting enables a focused view of a single category, reducing visual clutter (right).

The user can also chose between different types of coloring schemes for the paths: 1) color by the category at the first or last year (depending on the propagation direction), which shows the level of variation for a category over the length of the visualization; 2) color by transition, where the transition has a gradient from the source to target category color, useful for showing overall trends; and 3) color by reverse transition, where the transition path has a gradient from the target category to the source category, useful for category-level analysis of the distribution of source and target categories at a particular time step's category (Figure 3). To reduce visual clutter, there is also an option to look at only static transitions (i.e. no change in category between time steps), and to look at only variations (i.e. only changes in the categories).

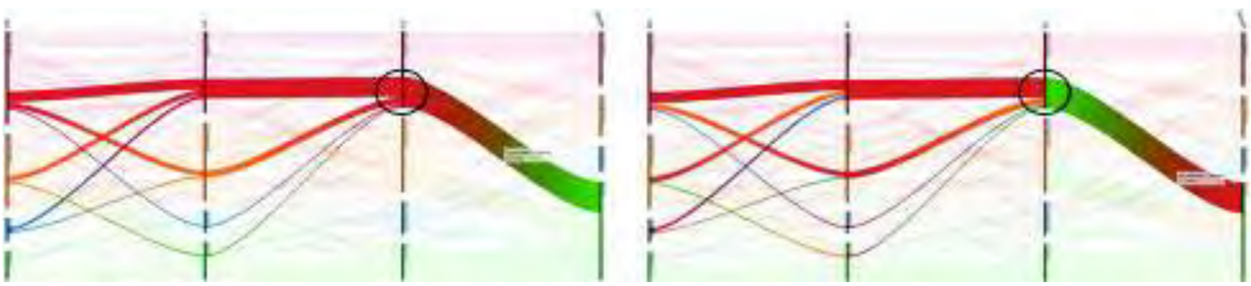


Figure 3. In addition to coloring by the starting category, paths can be colored by a gradient from source to target category (left), which redundantly encodes the category at each axis to emphasize overall trends, or by target to source category (right), which enables a rapid analysis of where paths are moving to/from at each category. The circled regions highlight this difference. On the right, it is immediately obvious what category this trajectory came from at death (Normal in green) and how this group is distributed at the previous time step.

Acknowledgments

This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation. We acknowledge the assistance from Meghana Ganapathiraju in helping to refine the visualization. Our implementation was adapted from the d3.parsets reusable chart by Jason Davies.

References:

1. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y. Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol. 2010;171(10):1090-1098
2. Bendix F, Kosara R, Hauser H. Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006;12(4):133-140.
3. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE TransVis Comput Graph, 2012;18(12):2659-2668.

Appendix E

Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates

David Borland[1], Vivian L. West[2], W. Ed Hammond[2]
[1]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[2]Duke Center for Health Informatics, Duke University, Durham, NC

Abstract

We present radial coordinates, a multivariate visualization technique based on parallel coordinates. The visualization contains a number of features driven by the needs of health-related data analysis, such as integrating categorical and numeric data, and comparing user-selected subpopulations via ribbon rendering. We illustrate the utility of radial coordinates by exploring primary care trust (PCT) and practice-level data from the United Kingdom's National Health Service, using three examples: lung cancer rates among PCTs, various cancer rates among only London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Introduction

With the ever-increasing size and number of health-related datasets, new analytical tools are becoming necessary to enable enhanced understanding of the vast amount of information contained within. Visualization leverages the power of the human visual system to reveal patterns and relationships in data by mapping the data to visually salient features.

One of the challenges for visualization of health-related data is the desire to incorporate data of many types (e.g. lab results, demographics, medications, vital signs, and genomic data) from various sources. We have developed a multivariate visualization technique, radial coordinates, that enables visual analysis of a wide range of health-related datasets and handles both numeric and categorical data (Figure 1). Radial coordinates facilitates the interactive exploration of datasets to reveal patterns in the data, discover relationships between variables, and compare user-defined subpopulations. In this manner we support the pursuit of hypothesis formations that can elicit further inquiry and lead to new knowledge.

An overview of an initial radial coordinates prototype applied to query data was given previously.[1] In this paper we provide a more in-depth description of the various features of a new implementation, which includes several new features, and discuss its application to primary care trust (PCT) and practice-level data from the National Health Service (NHS) in the United Kingdom (UK). We present three examples illustrating the use of radial coordinates to explore the NHS data: lung cancer rates among PCTs, a comparison of various cancer rates among London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Previous Work

Our visualization is based largely on parallel coordinates, a multivariate visualization technique which represents each dimension as a parallel axis, and each data entity as a line connecting the entity's value at each axis.[2,3] Non-parallel arrangements of axes have also been investigated.[4] Our radial coordinates arrangement differs in that the radial layout maintains a square aspect ratio even with many axes, and enables utilization of the space in the center of the radial layout. Parallel coordinates have been combined with various other visualization techniques[5-7], including direct integration of scatter plots.[8,9] In our visualization we include a scatter plot based on the first two principal components to enhance the ability to find clusters in high-dimensional data in an intuitive manner (Figure 1a). Future work will explore combinations with other techniques. We also incorporate chords representing the correlations between axes in a manner similar to Circos.[10] Extensions to parallel coordinates for incorporating categorical data include parallel sets[11] and hammock plots.[12] Both represent multiple data points as paths between axes, with the number of data points encoded as path width. Our curve spreading technique incorporates categorical and continuous data while still enabling the visualization of individual data points (Figure 2). Various techniques have been developed to combine multiple data points to enhance the understanding of large datasets[13,14] and observe clusters via edge bundling techniques.[15,16] Our ribbon rendering technique enables enhanced visualization of user-selected data points, including overlaying information of statistical data (median value and quartile ranges) of interest to the health-care community (Figure 1b). Axis ordering is an important element of parallel coordinates visualizations, as it is typically easier to notice relationships between variables with adjacent axes.[17-19] We employ a correlation-based clustering technique and also introduce dynamic reordering of categorical axis values to cluster similar values based on user-defined selections (Figures 3c, 3d).

Methods

Data

PCTs, abolished in 2013 due to NHS reorganization, were regional administrative bodies in the UK responsible for commissioning health services from providers and providing community health services. Here we investigate 26 variables measuring various health and socioeconomic factors for 147 of the 152 PCTs in England (five were removed due to missing data). Health factors include cancer rates, drug prescription rates, and factors related to diabetes prevalence and treatment. Socioeconomic factors include socioeconomic deprivation, economic output, geographic region, and local region classification (e.g. *Manufacturing Towns* and *Coastal and Countryside*) from the Office for National Statistics (ONS).

We also demonstrate our visualization with data showing the prevalence of a number of medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). There were ten SHAs in England from 2006-2013.

Visualization

The radial coordinates visualization, implemented using the D3 JavaScript library[20], represents each variable in a multivariate dataset by an axis, with the axes arranged radially around a circle. Each individual entity is represented by a curve that connects the value of that entity at each axis. Figure 1 gives an example applied to PCT data, with four PCT curves highlighted in red by the user.

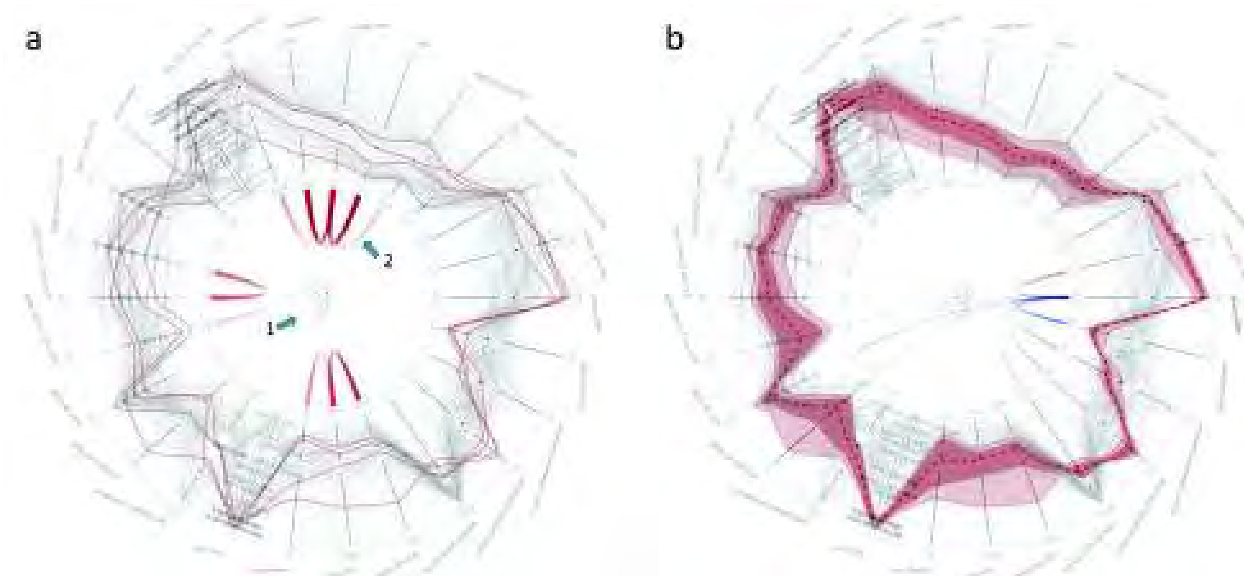


Figure 1. Radial coordinates visualizations of NHS PCT data. User-highlighted curves (red) enable the comparison of four PCTs across multiple variables (a). A linked scatterplot of the first two principal components can help show clusters in high-dimensions (a1). Chords connecting axes represent correlations (positive: red, negative: blue) above a user-defined threshold (a2). Ribbon rendering enables a simplified representation of user-defined subpopulations, displaying the data range optionally overlaid with median value and inner quartile ranges (b). Mouse over of an axis shows all correlations with that axis, regardless of user-defined threshold (b).

User selection of individual curves enables a visual comparison of how different entities relate across the various axes. A radial layout elegantly handles large numbers of axes while maintaining a square aspect ratio, also enabling the use of the center of the layout for supplemental visualizations, such as axis correlation chords and a scatterplot of the first two principal components (Figure 1a). Ribbon rendering uses a sliding window algorithm to draw the area between the innermost and outermost boundary of selected curves in a semi-transparent solid color, making it easier to see the spread of each subpopulation. An optional summary statistic overlay shows the inner quartile range and median value of each subpopulation (Figure 1b). Other visualization features include data-type dependent axis distribution visualizations and curve spreading for categorical and discrete data (Figure 2).

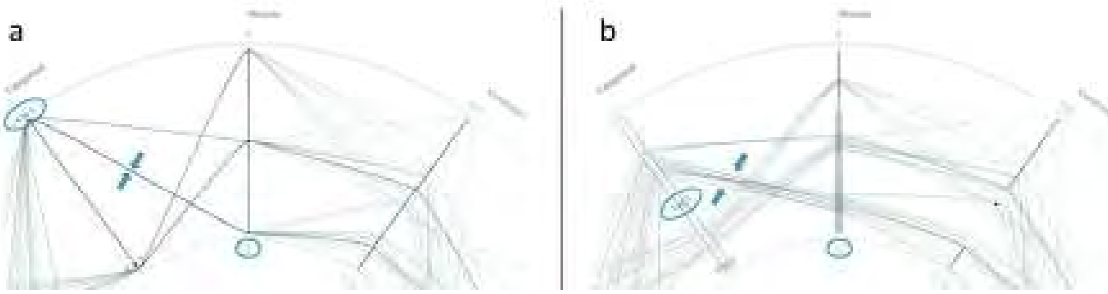


Figure 2. A sample data set without (a) and with (b) data-type dependent axis distribution visualizations and curve spreading. Axis distribution visualizations represent categorical axes as a stacked bar chart, discrete numeric axes as a histogram, and continuous numeric axes as a quartile plot[21], enabling rapid evaluation of the data type and overall distribution of the data for each axis. Curve spreading for categorical and discrete axes enables improved visualization of individual curves and clusters of curves, such as the number of data points with a Categorical value of Cat 1 and a Discrete value of three (highlighted in blue).

Results

Lung Cancer Prevalence

In Figure 3 the user has clicked on the lung cancer rate axis (*lung_Combined_DSR*), causing PCTs in the upper quartile of lung cancer rate to be automatically colored red, and the lower quartile blue. High and low lung cancer rates can now be compared across all dimensions in the data (Figure 3a). In the upper portion of the visualization it is apparent that PCTs with high and low lung cancer rates also tend to have high and low values for *extent*, *average_score*, *average_rank*, and *local_concentration* (also indicated by the correlation chords connecting these axes), which represent measures of social deprivation (poverty rate, socioeconomic status, etc.)

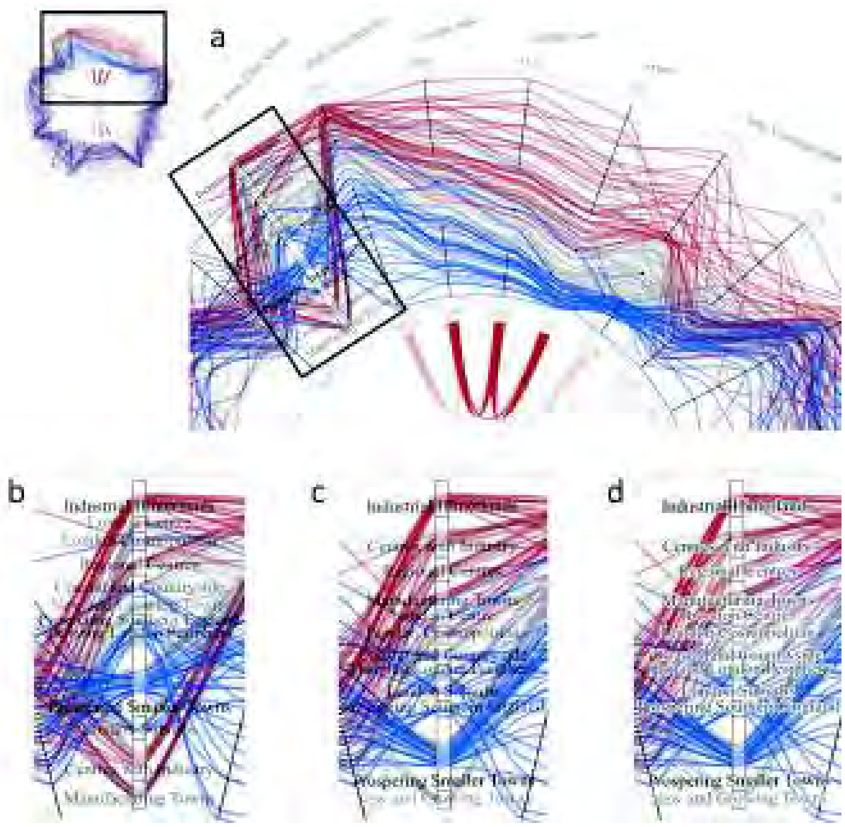


Figure 3. Visualization of lung cancer rates (red = upper quartile, blue = lower quartile) in 147 primary care trusts (PCTs) in the UK. High and low lung cancer rates tend to cluster based on regional classification (b), made clearer with automatic categorical axis reordering to cluster similar regions (c, d).

The red and blue curves also form clusters on the *ONS_Area_Class_Group* axis, a local region categorization from the ONS. Investigating this axis (Figures 3b-d) shows that *Industrial Hinterlands*, *Centres with Industry*, *Regional Centers*, and *Manufacturing Towns* all have high lung cancer rates, whereas *Prospering Smaller Towns*, *Prospering Southern England*, *London Suburbs*, and *Thriving London Periphery* all have low lung cancer rates. The discovery of such relationships via exploring the data visually drives the formation of causal hypotheses (e.g. pollution levels or smoking prevalence), which can be investigated further.

London Suburb Comparison

In Figure 4a a single PCT, Harrow, was seen to have the lowest lung, bladder, and colorectal cancer rates compared to all other PCTs, and has been highlighted in red. Harrow is classified as a *London Suburb*, so in Figure 4b the user has highlighted the other London suburbs in blue for comparison, made easier in Figure 4c via ribbon rendering. Harrow is shown to have a much higher value for the *oesophageal_Combined_SRR* axis, and thus a much higher esophageal cancer rate, than the other London suburbs, which are almost all in the lower quartile. This visualization raises the question of why Harrow has such a disparity in the rates of different cancers.

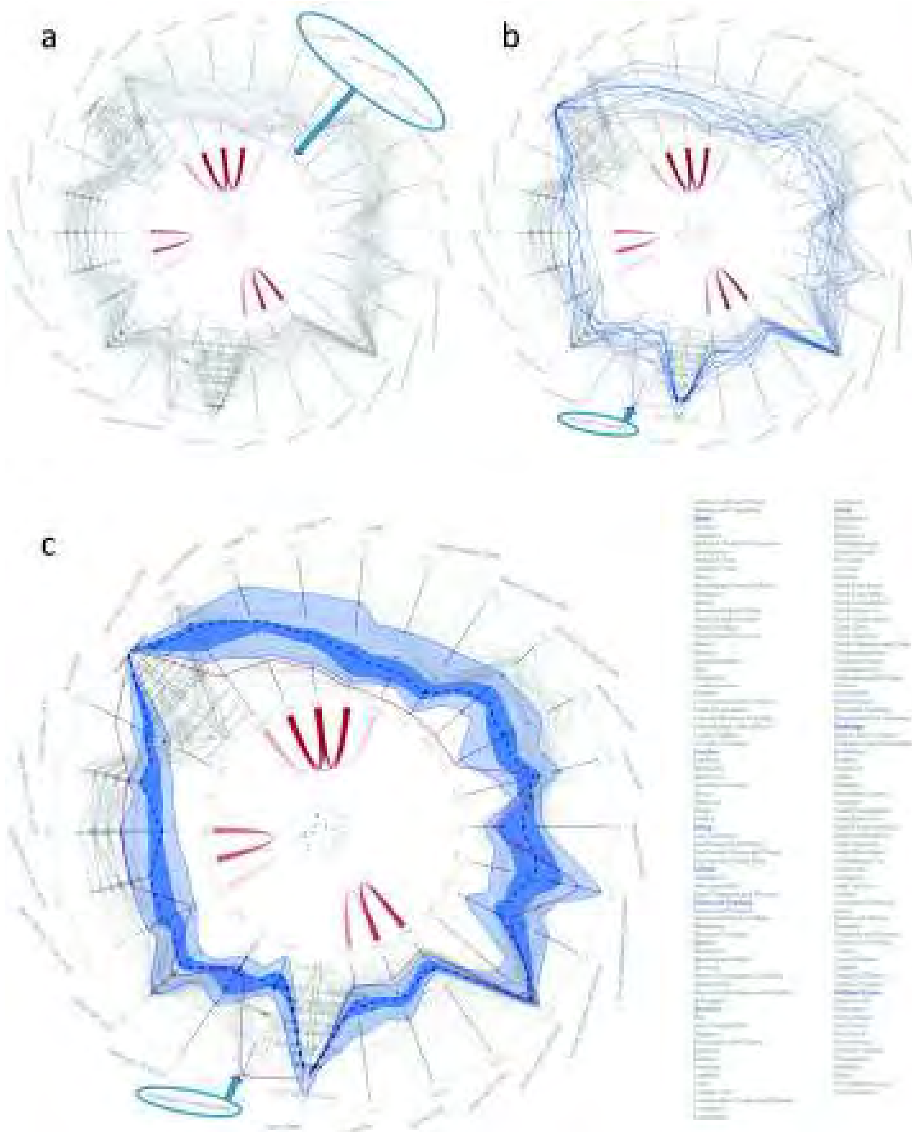


Figure 4. The Harrow PCT (red) has the lowest lung, bladder, and colorectal cancer rates (circled) among all 147 PCTs in the NHS dataset (a). Comparing Harrow to the other London suburbs (blue) reveals that Harrow has a much higher esophageal cancer rate (circled) than the other suburbs (b). Ribbon rendering makes it easier to visually compare Harrow with the other London suburbs (c).

According to the 2011 Census[22] Harrow is very diverse, with 63.8% of its population from the Black and Minority Ethnic communities, including the highest concentration of Sri Lankan Tamils and Gujarati Hindus in the UK and Ireland. India is known to have relatively low cancer rates in general, but some of the highest rates for oral and esophageal cancers in the world[23], which may help explain this phenomenon. Although further analysis is necessary, this example shows the utility of radial coordinates and ribbon rendering to compare subpopulations.

Practice-Level Data

Figures 8a and 8b show the prevalence of various medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). Figure 8a highlights in red two practices that appear to be outliers in the PCA scatterplot. Ribbon rendering makes apparent that they have the two highest prevalences for 12 of the 21 medical problems represented in the data. Figure 8b applies ribbon rendering to the remaining 1502 practices, making it easier to compare maximum and minimum values of medial problem rates for the two subpopulations.

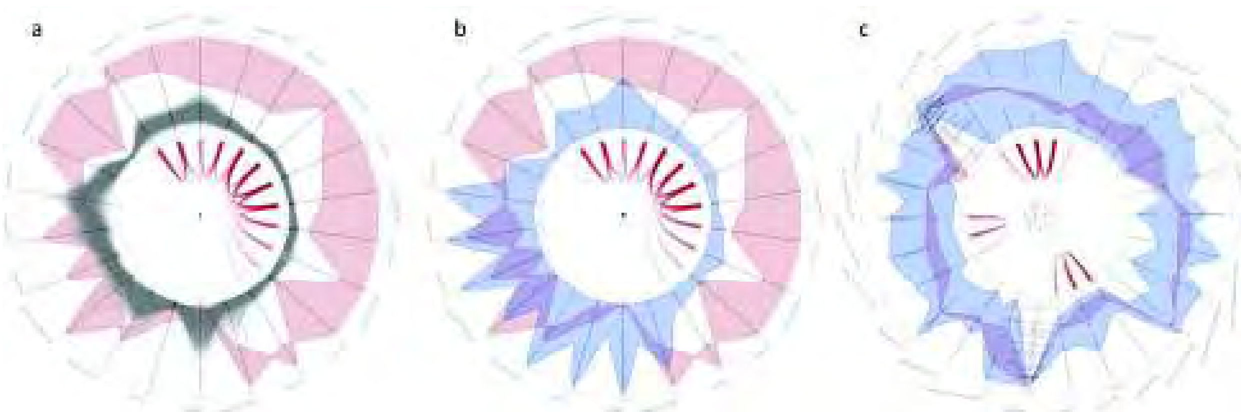


Figure 8. Two out of the 1504 practices in the London SHA, highlighted in red, have the two highest prevalences for 12 of the 21 medical problems represented in the NHS practice-level data (a and b). Comparing the PCTs containing these practices (red) to all other London PCTs (blue) does not reveal any major differences (c).

The two practices highlighted in red are Royal Hospital Chelsea in the Kensington and Chelsea PCT, and Nightingale House in the Wandsworth PCT. Because these two practices stood out so dramatically in the practice-level data, the user performed a PCT-level comparison of all London PCTs (Figure 8c). Interestingly, the Kensington and Chelsea and the Nightingale House PCTs (red) do not appear very different when compared to the other London PCTs (blue). Further research determined that Royal Hospital Chelsea is a retirement and nursing home for British soldiers and Nightingale House is a nursing home for the Jewish community that specializes in dementia, which may explain the high prevalence of problems such as dementia, hypertension, stroke, heart failure, and cancer in these two practices.

Conclusion

We have presented radial coordinates, a multivariate visualization technique based on parallel coordinates that incorporates features, such as per-axis population distribution visualizations based on data type (continuous, discrete, and categorical), direct visualization of correlations between variables, curve spreading for discrete and categorical data, visualization of summary statistics for user-selected subpopulations via ribbon rendering, and automatic reordering of categorical values based on user selection, driven by the needs of health-related data visualization.

We have applied radial coordinates to data from the UK's NHS at both the PCT and individual practice levels. Visualization of lung cancer rates among PCTs discovered possible relationships among lung cancer rate, socioeconomic factors, and regional classification. A comparison of London suburb PCTs revealed a potentially interesting PCT with a much higher esophageal cancer rate than other similar PCTs. Visualizing medical problem prevalence among over 1500 London practices showed two practices that have much higher rates of many medical problems. These examples illustrate the utility of the combination of visualization techniques embodied in our radial coordinates tool, and underline the need for further research in the use of visualization to aid in the analysis of complicated health-related datasets.

Acknowledgments

NHS and other UK data were made available courtesy of the BT Health Cloud. This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

References

1. West V, Borland D, Hammond WE. Visualization of EHR and Health Related Data for Information Discovery. In Proceedings of the 2013 AMIA Workshop on Visual Analytics in Healthcare. November 2013.
2. Gannet H. General summary, showing the rank of states, by ratios. 1880.
3. Inselberg A. The plane with parallel coordinates. Visual Computer. 1985;1(4):69-91.
4. Tominiski C, Schumann H. An event-based approach to visualization. In Proceedings of the Eighth International Conference on Information Visualization (IV'04). July 2004;101-107.
5. Rodrigues Jr. JF, Traina AJM, Traina Jr. C. Frequency plot and relevance plot to enhance visual data exploration. In Proceedings of the XVI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'03). 2003;117-124.
6. Edsall RM. The parallel coordinate plot in action: Design and use for geographic visualization. Computational Statistics and Data Analysis. 2003;43(4):605-619.
7. Siirtola H. Combining parallel coordinates with the reorderable matrix. In Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization. July 2003;63-74.
8. Holten D, van Wijk JJ. Evaluation of cluster identification performance for different PCP variants. Computer Graphics Forum. 2010;29(3):793-802.
9. Harter JM, Wu X, Alabi OS, Phadke M, Pinto L, Dougherty D, Petersen H, Bass S, Taylor II RM. Increasing the perceptual salience of relationships in parallel coordinate plots. In Proceedings of SPIE Visualization and Data Analysis 2012. January 2012.
10. Krzywinksi M, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Research. September 2009;19(9):1639-1645.
11. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics. July/August 2006;12(4):558-568.
12. Schonlau M. Visualizing categorical data arising in the health sciences using hammock plots. In Proceedings of the Section on Statistical Graphics, American Statistical Association. 2003.
13. Fua YH, Ward MRE. Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of the Conference on Visualization '99: Celebrating Ten Years. 1999;43-50.
14. Heinrich J, Weiskopf D. Continuous parallel coordinates. IEEE Transactions on Visualization and Computer Graphics. 2009;15(6):1531-1538.
15. Zhou H, Yuan X, Qu H, Cui W, Chen B. Visual clustering in parallel coordinates. Computer Graphics Forum. May 2008;27(3):1047-1054.
16. Heinrich J, Luo Y, Kirkpatrick AE, Zhange H, Weiskopf D. Evaluation of a bundling technique for parallel coordinates. In Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications. 2012;594–602.
17. Ankerst M, Berchtold S, Keim DA. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In Proceedings of the IEEE Symposium on Information Visualization. 1998;52-60.
18. Peng W, Ward MO, Rundensteiner EA. Clutter reduction in multi-dimensional data visualization using dimension reordering. In Proceedings of the IEEE Symposium on Information Visualization. 2004;89-96.
19. Seo J, Shneiderman B. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In Proceedings of the IEEE Symposium on Information Visualization. 2004;65-72.
20. Bostock M, Ogievetsky V, Heer J. D3: Data-driven documents. IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis). 2011;17(12).
21. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire, CN:Graphics Press. 2001.
22. Office for National Statistics. 2011 Census: Ethnic group, local authorities in England and Wales. 2012.
23. Sinha R, Anderson DE, McDonals SS, Greenwald P. Cancer risk and diet in India. Journal of Postgraduate Medicine. July-September 2003;49(3).

Appendix F

Temporal Visualization of Diabetes Mellitus via Hemoglobin A1c Levels

Eugenia McPeek Hinz[1], David Borland[2], Hina Shah[3], Vivian L. West[3], W. Ed Hammond[3]
[1]Duke Health Technology Solutions, Duke University; [2]RENCI, The University of North Carolina at Chapel Hill;
[3]Duke Center for Health Informatics, Duke University

Abstract

Diabetes mellitus is a chronic long-term disease requiring consistent medical treatment to achieve glucose control and prevent complications. Time of diabetes diagnosis can be variable and delayed years beyond disease onset. The spectrum of glycemic trajectories for a general population over an entire diabetes disease course is not well defined. Aligning disease course on death enables coherent data visualization. Our temporal visualization tool uses a parallel-sets inspired technique that illustrates the complicated and varied trajectories of hemoglobin A1c levels for a general diabetic population. A consistent glucose normalization trend for the majority of patients is seen over the course of their disease, especially in the six months prior to death. This tool permits discovery of population-level Hemoglobin A1c trends not otherwise evident without disease phase synchronization. These findings warrant further investigation and clinical correlation. Visualizations such as this could potentially be applied to other chronic diseases and spur further discoveries.

Introduction

Diabetes mellitus is a chronic disease that affects millions worldwide, resulting in numerous cardiovascular and renal complications, and subsequently is a major cause of death. Age of onset, duration of diabetes, and poor glycemic control are well-defined risk factors for the development of complications associated with increased mortality in persons with diabetes mellitus.[1] To decrease the development of complications associated with diabetes, tightly controlled glucose is the standard of care.[2] Notably some large prospective trials have found either worse outcomes or lack of benefit for some patients at high risk for complications under tight treatment control regimens.[3,4] Hemoglobin A1c (HbA1c), a marker of glucose control over the two to three months preceding the test, is a validated predictor of diabetes-related complications.[2] Using HbA1c to understand trajectories and temporal patterns of glycemic control over an entire diabetes disease course could be an important factor in improving treatment and reducing overall complications.

Data visualization techniques offer opportunities to explore large datasets and identify clinical patterns that might otherwise not be obvious. In this study we present a cohort of patients with diabetes (via ICD9 codes) from Duke University's data warehouse, visualizing their HbA1c levels over time, aligned by death, to explore trends of glycemic control. To the best of our knowledge, temporal visualization of glycemic control for a diabetic population standardized on death has not previously been presented. Our visualization groups HbA1c values into ordered categories of glycemic control (Normal, Borderline, Controlled, and Uncontrolled), utilizing a method based on parallel sets[5] and Sankey diagrams[6] to view temporal patterns in HbA1c values. We incorporate a number of features to facilitate interactive data exploration, such as viewing the progression of values either forwards or backwards in time, the ability to change the temporal sampling and range of the data being viewed, highlighting of multiple subpopulations, coloring based on the category along each path in the data or at the beginning/end of each path, and the incorporation of demographic data, such as gender.

Related Work

<u>Analysis of diabetes indicators</u>

A reduction in HbA1c levels lowers the risk of diabetes-related complications and mortality, especially for patients earlier in their disease course.[7] Counterintuitively, intensive treatment of glucose to reach near-normal levels for patients already experiencing diabetes-related complications has failed to lower all-cause mortality.[3] While large cross-sectional studies of populations such as the National Health and Nutrition Examination Survey find a temporal trend toward improving glycemic control over time, less well-established is the temporal trajectory of glycemic control for diabetic patients in general.[8] The only other work the authors are aware of looking specifically at glycemic control trajectories for a large diabetic cohort followed patients prospectively to the end point of death.[9] The study correlated initial glucose control to outcome of death, but did not report specifically on the population glucose trajectories.

Visualization methods

Our visualization tool is based on parallel sets[5] and Sankey diagrams.[6] Parallel sets were originally developed for visualizing relationships in multivariate categorical data, whereas Sankey diagrams, introduced by M. H. P. R. Sankey, are typically used for describing the flow of quantities such as energy, material, or cost. The original parallel sets user interface enables user-defined classification definitions, statistical analysis information, and various sorting methods. Parallel sets combines the concepts of parallel coordinates[10] and mosaic plots[11], enabling an aggregation of data points within visualization elements, as opposed to showing each individual element, which is typical of parallel coordinates. Multiple systems aggregate data points for summary.[5,12-14] For example, EventFlow enables the search and visualization of interval data, such as periods of medication treatment, to examine the order of sequences of events in the data.[12] OutFlow facilitates analyses of temporal event data in the form of pathways with relevant statistics.[14] All of these visualization tools look at event occurrences and their order, without placing these events on time axes. Our diabetes visualization uses the parallel sets paradigm, with each axis representing a temporal sample of HbA1c levels instead of a separate variable, similar to von Landesberger et al.[13] Although our current dataset is relatively small (121 patients), we chose a parallel sets representation in part due to its ability to aggregate many data points. The visual complexity is bounded by the number of axes and categories per axis, not by the number of data points, making it suitable for the exploration of larger datasets in the future. This representation also easily incorporates additional non-temporal variables, such as demographic data.

Methods

Data extraction and preprocessing

Data from Duke University's data warehouse were extracted using DEDUCE, an on-line query tool developed at Duke to assist researchers in human subjects research and departments seeking quality improvement data.[15] Beginning with over 4.4 million patients, we first queried by 23 IDC9 codes for diabetes mellitus, with and without complications. The query was further refined by querying on patient death indicator and laboratory tests for glycosylated hemoglobin (HbA1c), and finally by including only patients prescribed anti-hyperglycemics. This search returned data from 208 patients. From this cohort of 208, we eliminated four that did not have a year of death recorded, one whose date of death was documented but continued to have laboratory results recorded after that date, and 82 who did not have at least 10 years of HbA1c laboratory values. Our final cohort includes data from 121 patients.

We average HbA1c values, given as a percentage of total hemoglobin, over 6 month time intervals. In the case of missing HbA1c values within a 6 month period we first attempt to impute an HbA1c value from the average glucose (AG) values over that period of time, via the formula HbA1c = (AG + 46.7) / 28.7.[16] If no glucose values exist in that time period, the previous HbA1c value (measured or imputed) is carried forward. HbA1c values are then classified into four categories based on the severity of diabetes: Normal < 5.7, Borderline $[5.7, 6.5)$, Controlled $[6.5, 8)$, and Uncontrolled ≥ 8.

The sampled data is time-aligned by the death event for each patient. The visual representation of diabetes progression propagates backwards in time initially. Time is represented as number of years before death.

Visualization

Our visualization tool was developed using the D3 Javascript library.[17] The aim of this visualization is to investigate temporal trajectories of HbA1c levels for a large cohort of diabetes patients over a number of years prior to death. Since parallel sets is effective for showing relations between categories using aggregated frequencies of paths through categories at each dimension, it is a reasonable choice for showing HbA1c summary trajectories. The visualization tool shows a total of five categories: four representing glycemic control, and one optional Missing category for patients with data greater than 10 years before death. . Each vertical axis is a time step. The user can choose the frequency of these time steps, with a minimum sampling frequency of six months. The user can also select the maximum number of years before death.

The death event axis is placed at the right with all other time steps moving backwards in time to the left (Figure 1). Each vertical axis is split into the four HgA1c categories (Normal in green, Borderline in blue, Controlled in orange, and Uncontrolled in red), and a Missing category in grey prior to 10 years before death. The height of each axis category represents the proportion of the patients in that category at that point in time. Paths moving between axes recursively split moving backwards from death to show the trajectories of similar groups of patients. The visualization can show trends either starting at the death event i.e. going backwards in time (dividing

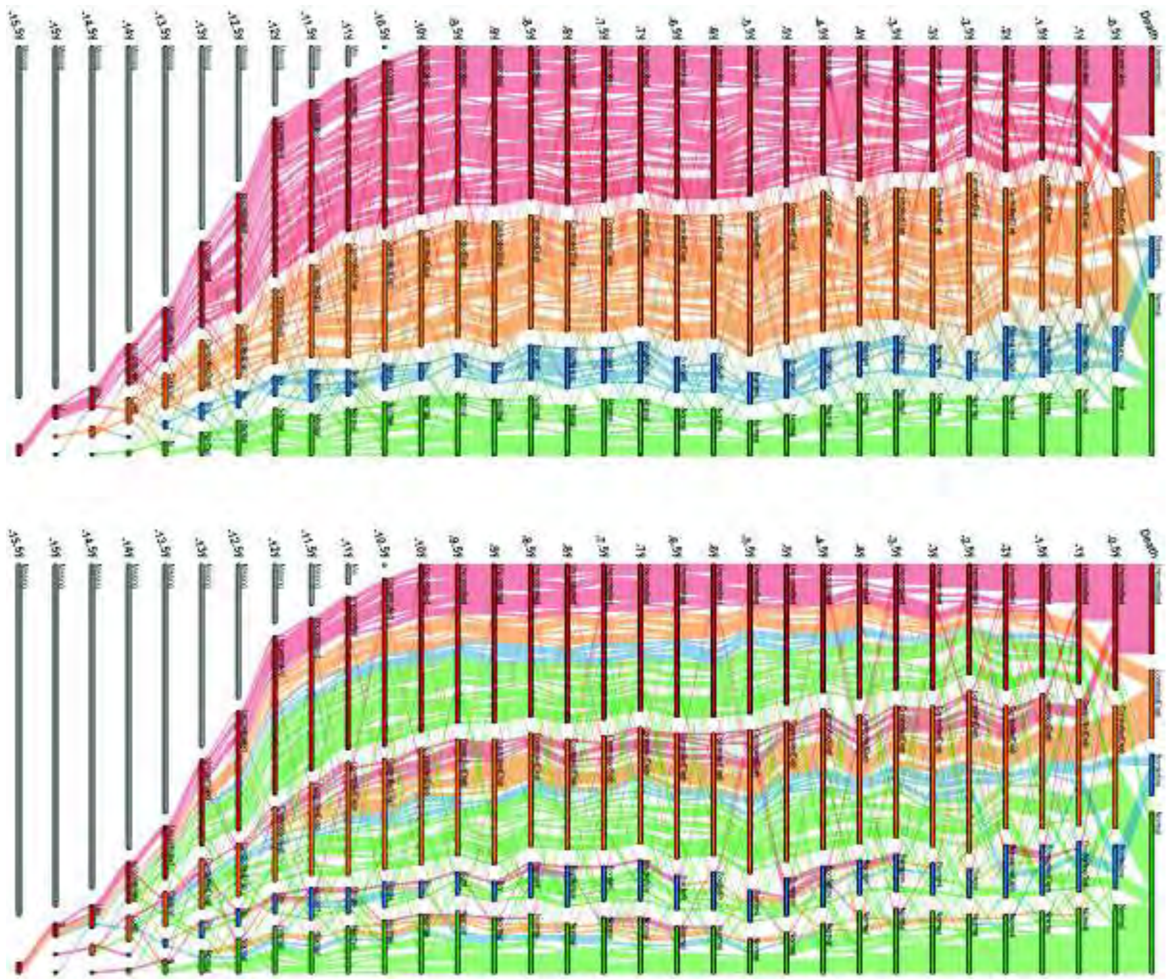


Figure 1. Diabetes progression overview visualizations. The top image colors paths by the current HbA1c at each time step, which is useful for emphasizing overall temporal trends. The bottom images colors paths by the HbA1c level at death, showing at each time step where each path will end.

recursively right to left), or starting at the last year in the visualization, i.e. going forward in time (recursive division from left to right). Going backwards and coloring by death shows at any time point the relationships between patients in a given category and their categories at death, while going forward in time shows the relationship between patients in a given category and their categories at a user-defined earlier point in time (Figure 2). Following Shneiderman's Mantra[18] of first overviewing and then filtering, the user can highlight one or more groups of patients by clicking on categories or trajectories to highlight the behavior of that group of patients going backward and forward in time, reducing visual clutter (Figure 2). A tooltip also shows the actual number of patients in each group and their percentage of the total population.

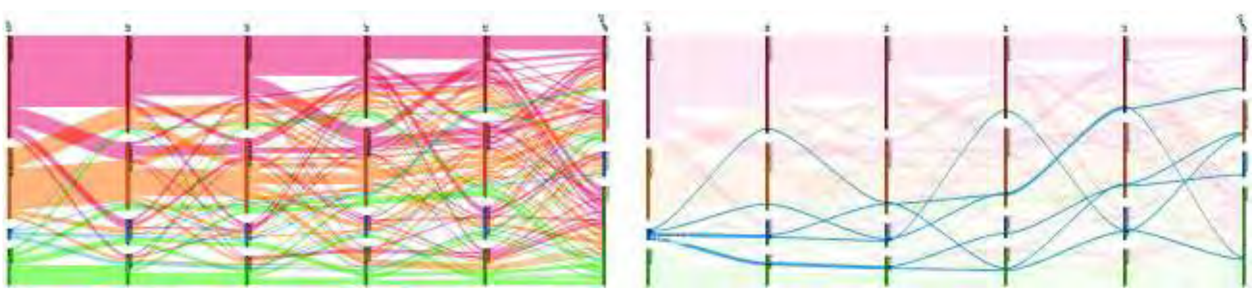


Figure 2. A 10-year range of data, sampled every two years, with forward propagation to show how the trajectories of patients change moving forward in time (left). Highlighting enables a focused view of a single category, reducing visual clutter (right).

The user can also chose between different types of coloring schemes for the paths: 1) color by the category at the first or last year (depending on the propagation direction), which shows the level of variation for a category over the length of the visualization, 2) color by transition, where the transition has a gradient from the source to target category color, which is useful for showing overall trends, and 3) color by reverse transition, where the transition path has a gradient from the target category to the source category, which is useful for category-level analysis of the distribution of source and target categories at a particular time step's category (Figure 3). To reduce visual clutter there is also an option to look at only static transitions (i.e. no change in category between time steps), and to look at only variations (i.e. only changes in the categories).

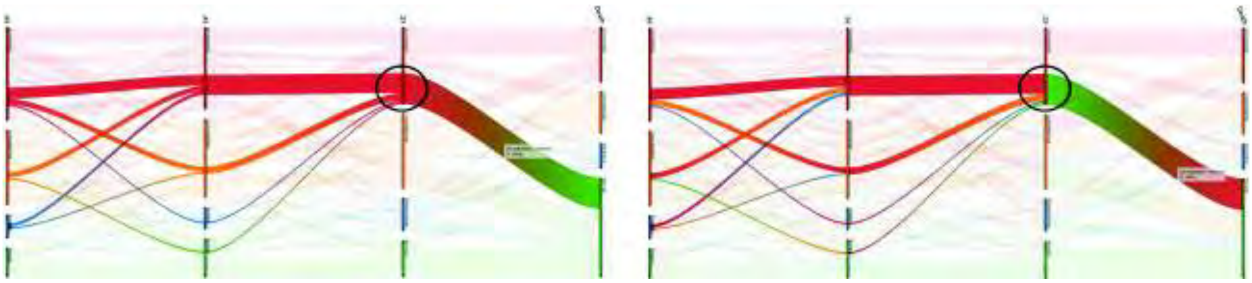


Figure 3. The user can observe separate groups by selecting individual trajectories. In addition to coloring by the starting category, paths can be colored by a gradient from source to target category (left), which redundantly encodes the category at each axis to emphasize overall trends, or by target to source category (right), which enables a rapid analysis of where paths are moving to/from at each category. The circled regions highlight this difference. On the right, it is immediately obvious what category this trajectory came from at death (Normal in green) and how this group is distributed at the previous time step.

We also include the ability to incorporate demographic data, such as gender, as additional axes (Figure 4). This feature enables the comparison of trajectories for different subpopulations based on data other than just HbA1c levels.

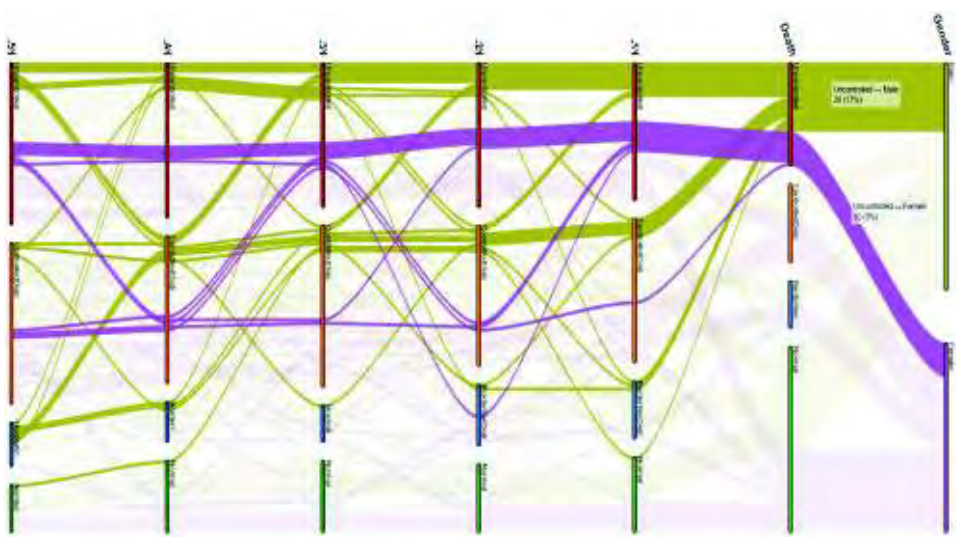


Figure 4. By adding a gender axis and selecting two groups we can compare the variability of males who were Uncontrolled at death (olive) to women who were Uncontrolled at death (purple). Men appear to have more variability over the 5-year period being visualized, as shown by the large number of transitions between different categories.

Findings

In the 10 years before death, there is a consolidating trend to improved glucose control across all diabetes control categories from uncontrolled to normal. Overall diabetes control shifts from uncontrolled diabetes for 46% of the cohort to 25% at death utilizing HbA1c and imputed glucose values. A reciprocal increase in combined borderline and normal range glucose control goes from 25% at 10 years out to 57% at death (Table 1). The trend for better glucose control is most visible in the last six months before death. The overall final glycemic trajectory is also evident in the bottom image from Figure 1 where the control category at death is colored retrospectively. Notably a small minority of the uncontrolled sub-group remains poorly controlled over the entire disease course. By including category temporal transitions this visualization also illustrates the complexity of the underlying data, with many trajectories exhibiting a large degree of variation in HbA1c categorization over time.

Table 1. Percent of patients by Diabetes control category over 10 years prior to death using HbA1c with imputed glucose results.

Glycemic Control by HbA1c	*-10 years years to death*	*-5 years years to death*	*At Death*
Uncontrolled Diabetes	46 %	39 %	25 %
Controlled Diabetes	32 %	39 %	19 %
Borderline	5 %	11 %	12 %
Normal	17 %	12 %	45 %

Discussion

The progression of diabetes with accumulating end organ complications is well recognized. There is a clinical presumption that diabetes-related complications are also associated with worsening glycemic control for patients with end stage diabetes mellitus. Since most prospective cohort studies are organized by a patient's clinical presentation, treatment or demographics, they tend to be cross sectional studies of a population and include patients across a disease continuum. By creating a cohort organized by a death criterion with 10 or more years of diabetes lab data, we have sub-selected a general but presumably more ill diabetic population. Phasing HbA1c values by death allows data coherence that translates into the visualization of glycemic trajectories that would be less evident in cross sectional studies of diabetic patients. Understanding the course of diabetes control is important to discerning differences in outcomes, treatments and identifying sub-phenotype populations.

Death event as an organizing point for temporal data visualization permits a clear starting point to observe the course of medically treated diabetes. Cause of death is not defined, so further characterization of subpopulations visualized in the cohort, like the always uncontrolled diabetes subgroup, warrants further clinical investigation to see if they are representative of the cohort overall. All patients in this cohort had data for at least 10 years, as such our population is specific for patients under some manner of regular medical care, and interpretation of the data with respect to populations with less regular medical care should be limited. Using the imputed average glucose and average HbA1c values aligned on the cohort's endpoint enables capture of all glycemic values, including those potentially before even the diagnosis of diabetes is made.

We observed a trend to normalization of HbA1c in the last year of life. The reasons behind improved diabetes control near the end of life could include multiple factors, such as increased insulin half-life due to impaired renal and hepatic metabolism, decreased dietary intake related to anorexia or nausea, and falsely low HbA1c secondary to uremia or anemia.[19] Interestingly, the goals for end-of-life treatment in diabetic patients are generally to limit side effects of either hyper or hypoglycemia and often entail a scaling back of treatment which would be expected to be associated with more hyperglycemia not less. By using visualization tools to see the progression of HbA1c values in diabetic patients in the years before their death, our findings of glucose normalization in light of this paradigm highlight the need for further clinical investigation and interpretation.

Our data visualization tool displays temporal patterns of diabetes metric across a population and for the last years of this disease continuum. Tools such as these can only display patterns that can potentially illuminate findings that need further clinical validation and statistical investigation to determine clinical significance if any.

Future work

The visualizations we have shown here represent a small number of patients in the dataset. This has enabled us to test and refine the visualization before using large amounts of data. Next we will include diabetes-related co-morbidities, e.g. cardiovascular, neurological, and renal manifestations of prolonged diabetes illness, and additional demographic variables, e.g. age and ethnicity. We plan to link this temporal visualization to other multivariate visualizations highlighting selected groups of patients, helping to show factors related to diabetes. We are also working toward a better statistical analysis of the data, and its representation in this tool. In particular, we wish to incorporate information regarding the amount of imputed and extrapolated data in the visualization.

Conclusion

Exploring the natural disease course of diabetes control with data visualization tools permits identification of potentially clinically important trends that would be difficult to recognize otherwise. Further investigation and definition on the clinical significance of the normalization of HbA1c in the final years of life are warranted.

Acknowledgments
This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation. We acknowledge the assistance from Meghana Ganapathiraju in helping to refine the visualization. Our implementation was adapted from the d3.parsets reusable chart by Jason Davies. We also acknowledge the assistance of Mark Massing MD PhD MPH and Susan Spratt MD for comments and discussions on graphical representations of diabetes control.

References

1. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med.1993;329(14):977-986.
2. American Diabetes Association. Standards of Medical Care in Diabetes. Diabetes Care. 2009;32(S1):S13-S61.
3. The Accord Study Group. Long-term effects of intensive glucose lowering on cardiovascular outcomes. N Engl J Med. 2011;364(9):818-828.
4. The UK Prospective Diabetes Study Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet. 1998;352(9131):837-853.
5. Bendix F, Kosara R, Hauser H. Parallel sets: visual analysis of categorical data. IEEE Symp on Info Vis. 2006;12(4):133-140.
6. Riehmann P, Hanfler M, Froehlich B. Interactive Sankey diagrams. IEEE Symp on Info Vis. 2005;233-240.
7. Holman R, Paul S, Bethel M, Matthews D, Neil H. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med. 2008;359(15):1577-1589.
8. Ford E, Li C, Little R, Mokdad A. Trends in A1C concentrations among U.S. adults with diagnosed diabetes from 1999 to 2004. Diabetes Care. 2008;31(1):102-104.
9. Gebregziabher M, Egede LE, Lynch CP, Echols C, Zhao Y. Effect of trajectories of glycemic control on mortality in type 2 diabetes: a semiparametric joint modeling approach. Am J Epidemiol. 2010;171(10):1090-1098.
10. Inselberg A, Dimsdale B. Parallel coordinates. Human-Machine Interactive Systems. 1991;199-233.
11. Hoffman H. Exploring categorical data: Interactive mosaic plots. Metrika. 2000;51(1):11-26.
12. Monroe M, Wongsuphasawat K, Plaisant C, Shneiderman B, Millstein J, Gold S. Exploring point and interval event patterns: Display methods and interactive visual query. HCIL Tech Report, University of Maryland. 2012.
13. von Landesberger T, Bremm S, Andrienko N, Andrienko G, Tekusova M. Visual analytics methods for categoric spatio-temporal data. IEEE Conf on Vis Anal Sci and Tech(VAST) 2012;183(192):14-19.
14. Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE TransVis Comput Graph, 2012;18(12):2659-2668.
15. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE guided query tool: Providing simplified access to clinical data for research and quality improvement. J Biomed Info. 2011;2:266-276.
16. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C Assay into estimated average glucose values. Diabetes Care. 2008;31(8):1473-1478.
17. Bostock M, Ogievetsky V, Heer J. D^3: Data-driven documents. IEEE Trans Visualization & Comp Graphics. 2011;17(2):2301-2309.
18. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualization. Proc. 1996 IEEE Symp Vis Lang. 1996;336-343.
19. Kalantar-Zadeh K, Derose SF, Nicholas S, Benner D, Sharma K, Kovesdy CP. Burnt-out diabetes: Impact of chronic kidney disease progression on the natural course of diabetes mellitus. J Renal Nutrition. 2009:19(1):33-37.

Appendix G

Exploring Novel Visualizations of Electronic Health Record Data

Meghana Ganapathiraju[1]
Mentors: David Borland, PhD[2], Vivian West, PhD[3], W.Ed Hammond, PhD[3]
[1]Green Hope High School, [2]RENCI, [3]Duke University

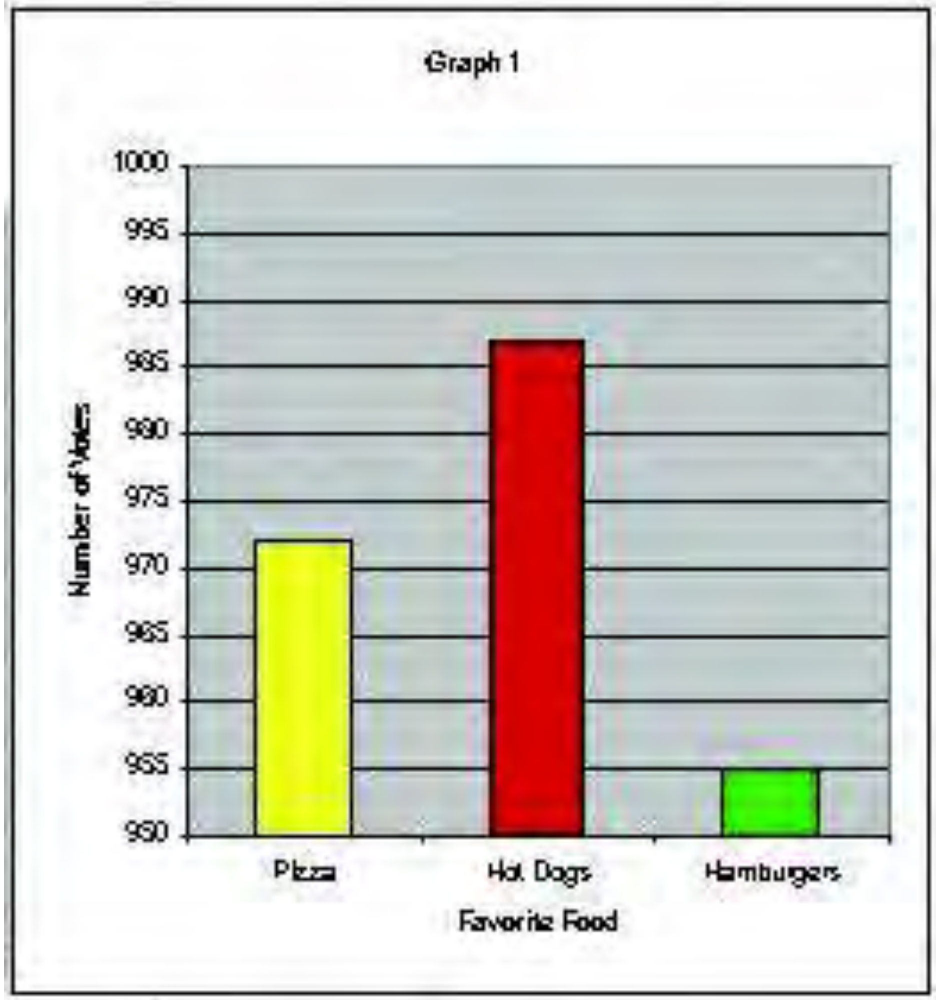
Information Visualization

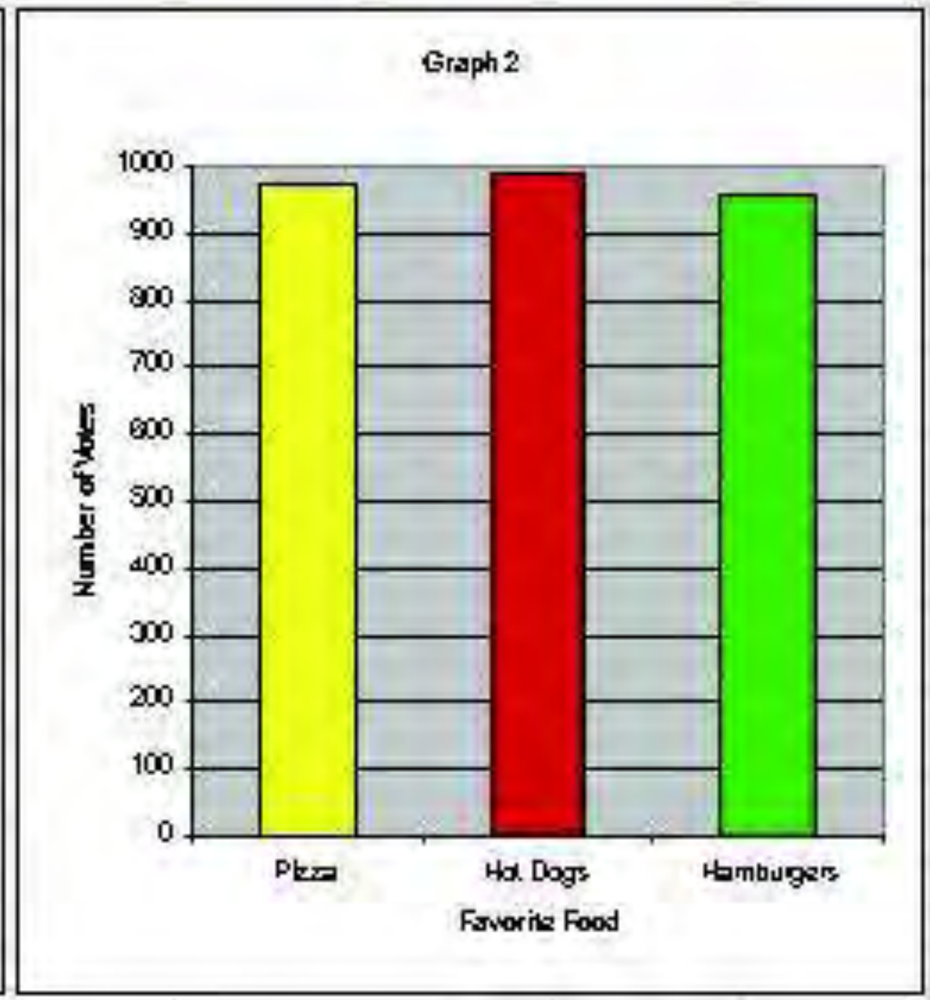
- The use of interactive visual representations of abstract data to amplify cognition (Ware, 2004)
- Makes interpretation of information easier
- Must correctly represent information or the visualization can be misleading or confusing
- Visualization can help detect key patterns otherwise difficult to pick out

John Snow’s Cholera Map

http://flowingdata.com/2007/09/12/john-snows-famous-cholera-map/

Information Visualization





Visualizing Multivariate Data

- Multivariate data has more than 2 or 3 dimensions
- Lose information if shown in traditional graphs/plots
- New methods of visualization are necessary

Multivariate visualization techniques

- Scatterplot Matrix
- Starplots
- Parallel sets/coordinates

Purpose

- Visualize EHR data
  - Systems are increasing in size and complexity
  - Researchers se the database in different ways

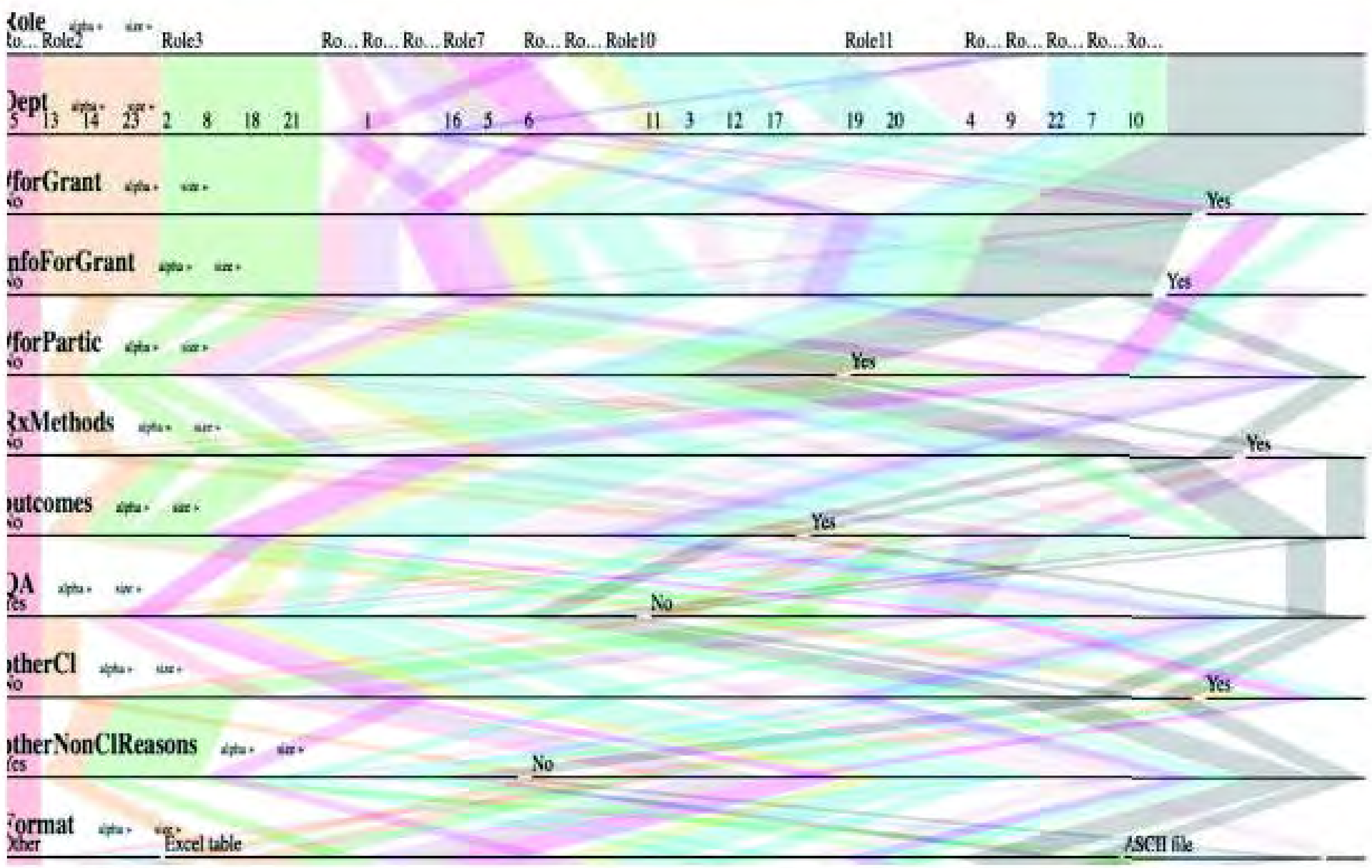
- Visual representations may facilitate additional insight

Methods

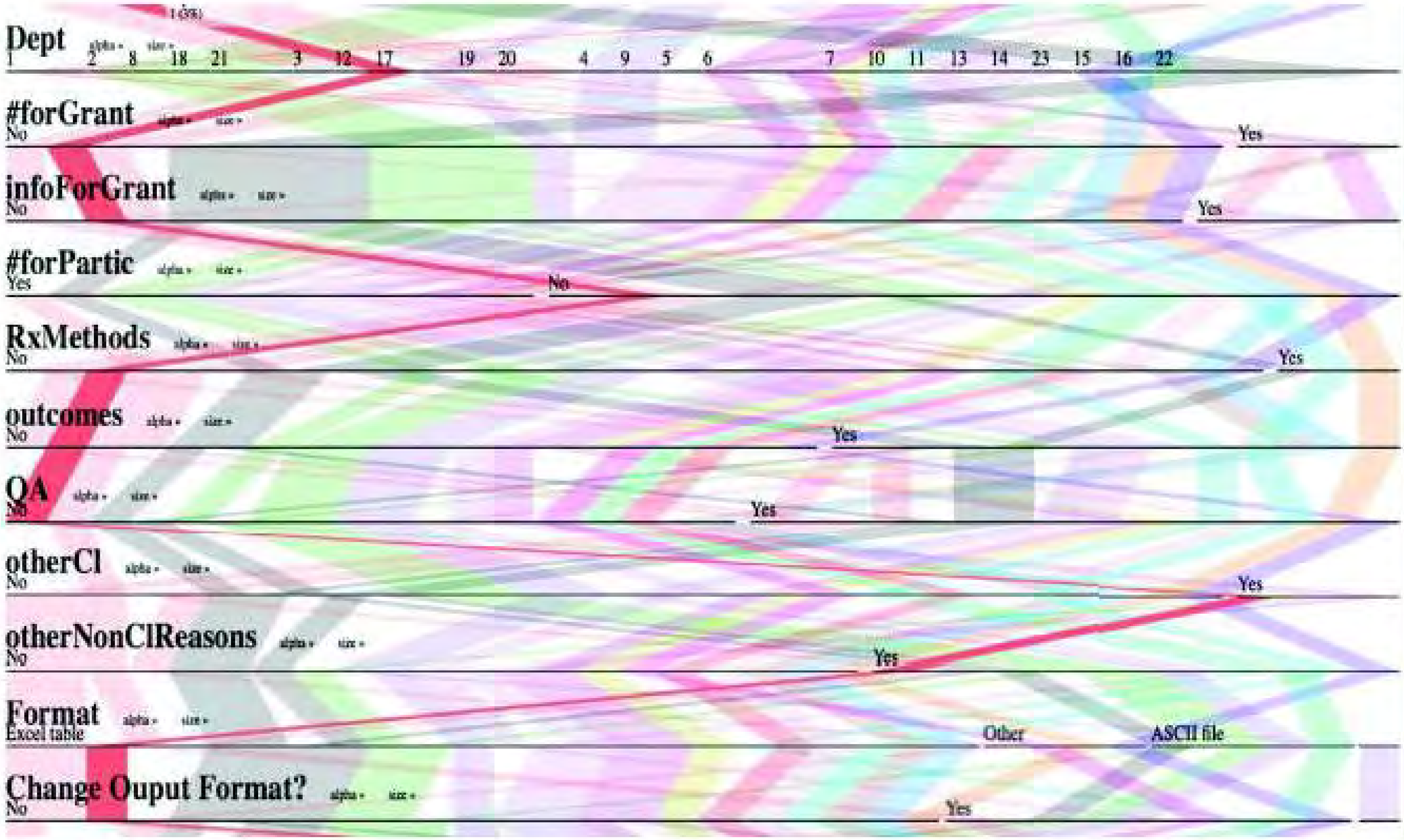
- Surveyed researchers
- Data stored as an excel file
- D3 JavaScript library used to produce the visualization

Parallel Sets Visualization

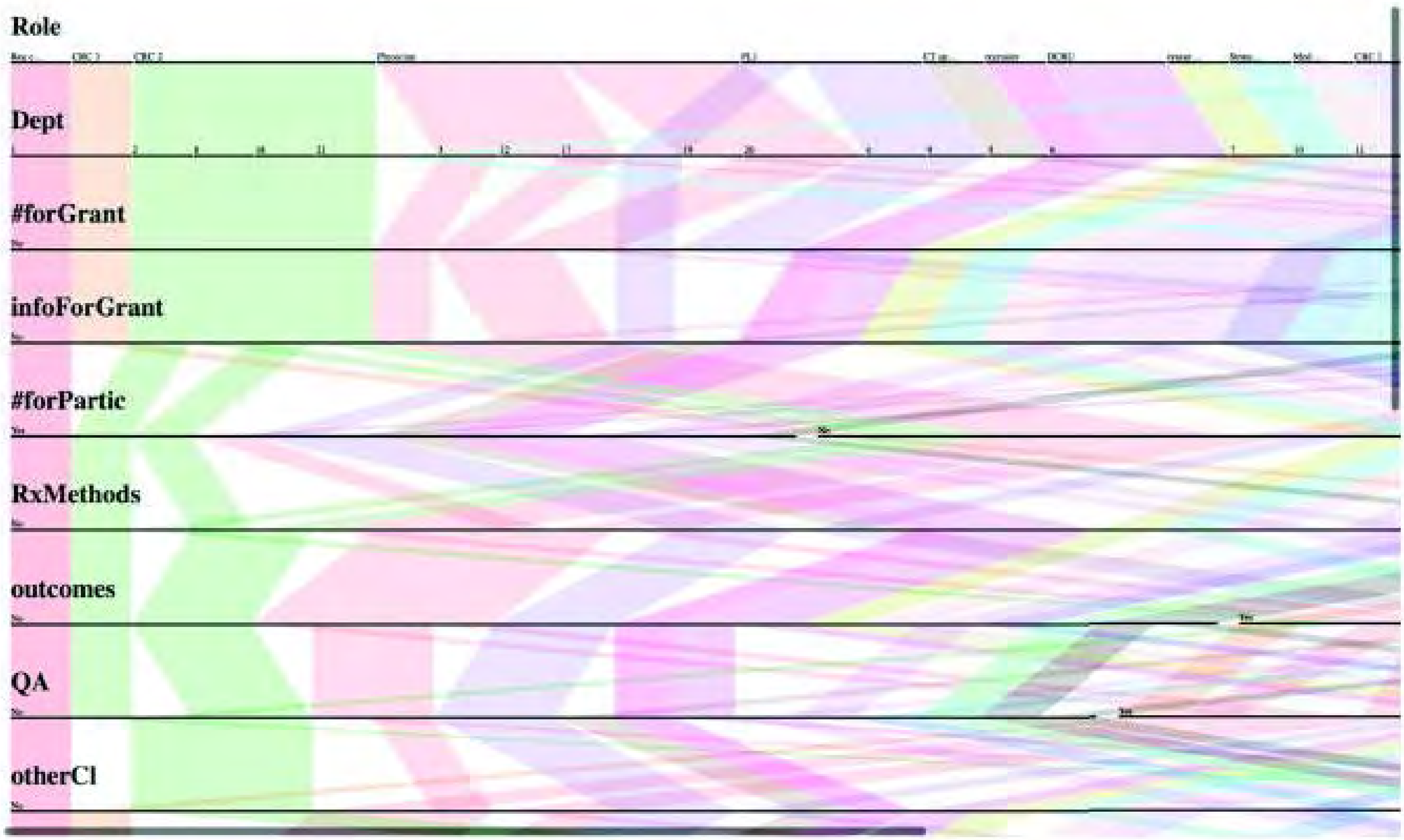
[Live Demo Link]



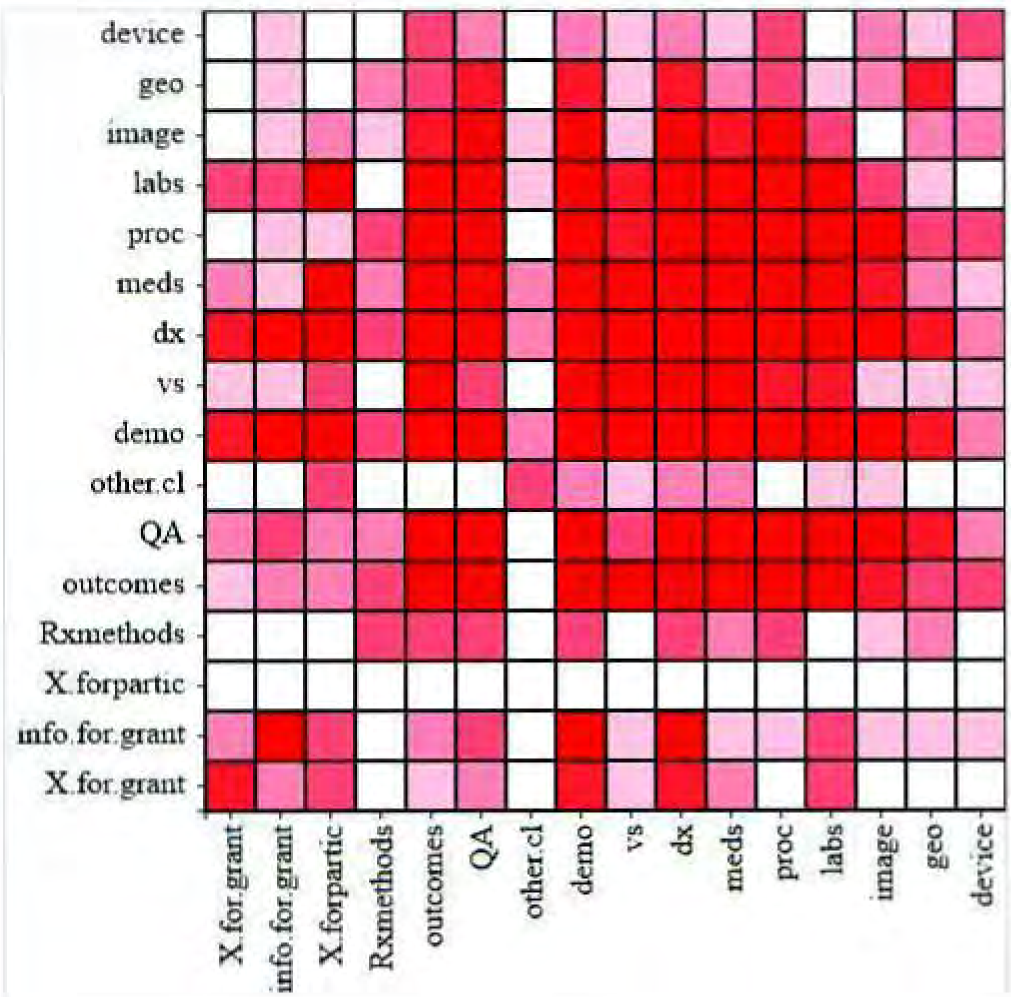
Parallel Sets Visualization



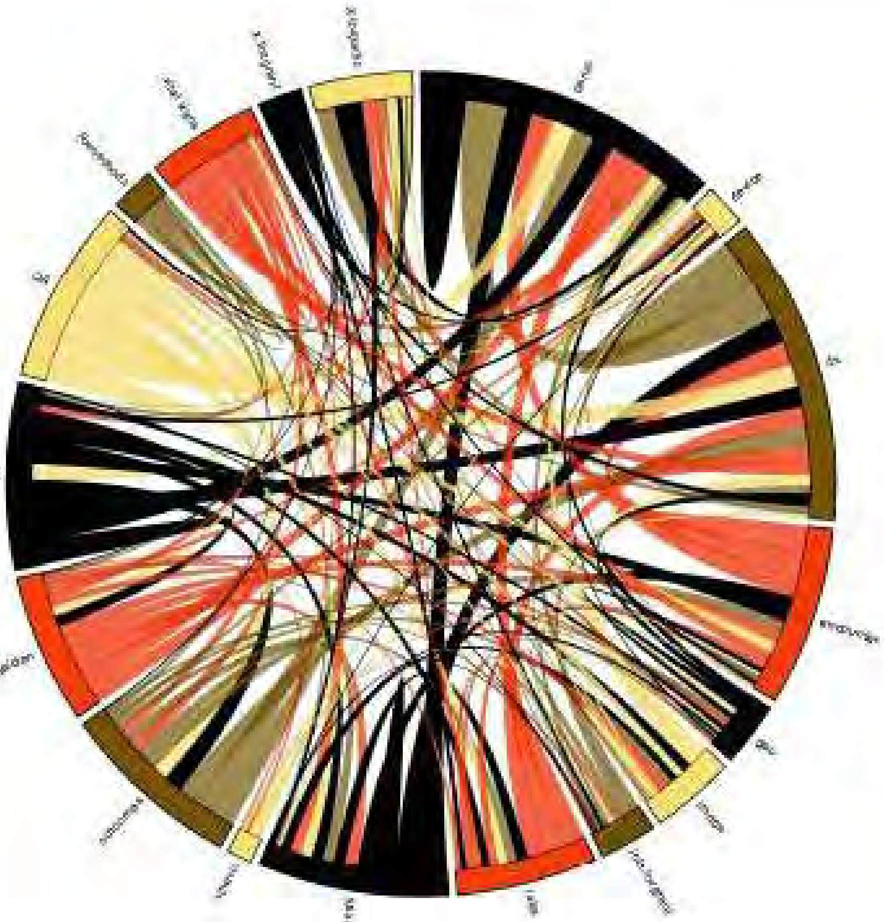
Parallel Sets Visualization

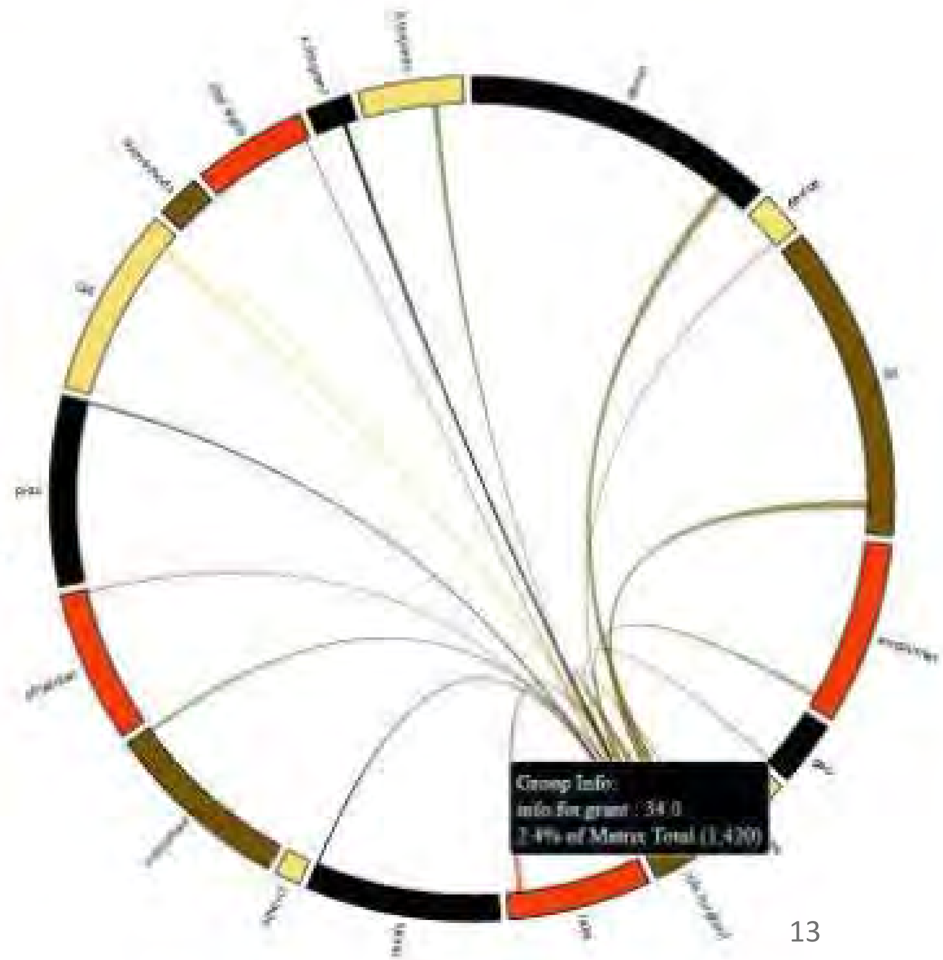


Co-occurrence Matrix



Chord Diagram





Future Work

- These visualizations are just some of the possible methods of multivariate visualization
- Explore further options
- Show visualizations to a variety of users

Acknowledgements

Dr. David Borland, RENCI

Dr. Vivian West, Duke University

Hina Shah, UNC Chapel Hill

Dr. Ed Hammond, Duke University

Acknowledgements

This work was supported by the US Army Medical Research and Material Command (USAMRMC) under Grant No. W81XWH-13-1-0061.

Thank You!

Appendix H

OPEN ACCESS

Innovative information visualization of electronic health record data: a systematic review

Vivian L West,[1] David Borland,[2] W Ed Hammond[1]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2014-002955).

[1]Duke Center for Health Informatics, Duke University, Durham, North Carolina, USA
[2]The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence to
Dr Vivian L West, Duke Center for Health Informatics, Duke University, 2424 Erwin Road, Suite 9002, Room 9021, Durham, NC 27705, USA; vivian.west@duke.edu

Received 5 May 2014
Revised 25 July 2014
Accepted 14 September 2014

ABSTRACT

Objective This study investigates the use of visualization techniques reported between 1996 and 2013 and evaluates innovative approaches to information visualization of electronic health record (EHR) data for knowledge discovery.

Methods An electronic literature search was conducted May–July 2013 using MEDLINE and Web of Knowledge, supplemented by citation searching, gray literature searching, and reference list reviews. General search terms were used to assure a comprehensive document search.

Results Beginning with 891 articles, the number of articles was reduced by eliminating 191 duplicates. A matrix was developed for categorizing all abstracts and to assist with determining those to be excluded for review. Eighteen articles were included in the final analysis.

Discussion Several visualization techniques have been extensively researched. The most mature system is LifeLines and its applications as LifeLines2, EventFlow, and LifeFlow. Initially, research focused on records from a single patient and visualization of the complex data related to one patient. Since 2010, the techniques under investigation are for use with large numbers of patient records and events. Most are linear and allow interaction through scaling and zooming to resize. Color, density, and filter techniques are commonly used for visualization.

Conclusions With the burgeoning increase in the amount of electronic healthcare data, the potential for knowledge discovery is significant if data are managed in innovative and effective ways. We identify challenges discovered by previous EHR visualization research, which will help researchers who seek to design and improve visualization techniques.

BACKGROUND AND SIGNIFICANCE

In 2004 a presidential executive order, 'Electronic Health Records (EHRs) for All Americans', laid out tenets to improve the quality and efficiency of healthcare, with one goal being accessible EHRs for most Americans within 10 years.[1 2] In September 2009, years of research and policy work culminated in the Health Information Technology for Economic and Clinical Health Act (HITECH Act) allocating $19.2 billion in incentives to increase the use of EHRs by hospitals and health delivery practices. The latest report from the Centers for Medicare and Medicaid Services (CMS) found that approximately 80% of eligible hospitals and over 50% of eligible professionals had received incentive payments from CMS for adopting EHRs.[3]

With the burgeoning amount of electronic data, the potential for knowledge discovery is significant if the large amounts of data are managed in innovative and effective ways. This review investigates data visualization techniques reported in the healthcare literature between 1996 and 2013, aiming to evaluate innovation in approaches to information visualization of EHR data for knowledge discovery.

Historical background

The graphical visualization of data dates back to the later part of the 18th century when William Playfair is credited with the first use of line graphs, pie charts, and bar graphs. Playfair, an engineer and economist, considered charts and graphs the most effective way to communicate information about data.[4] In 1786, he published *The statistical breviary; shewing, on a principle entirely new, the resources of every state and kingdom in Europe; illustrated with stained copper plate charts, representing the physical powers of each distinct nation with ease and perspicuity*, stating that by graphically representing data, the reader can best understand and retain the information.[5]

A widely recognized visualization is a two-dimensional graph using time, temperature, and geography showing Napoleon's march on Russia in 1812. This linear graph, published by Charles Minard in 1861, shows the movement of Napoleon's army across Russia to Moscow and back to Europe (figure 1).

The horizontal axis of the graph is marked with temperatures below freezing as they returned. The width of the line depicting the army is scaled, illustrating the decline in the number of men returning from war as the temperature decreased, which can be easily compared to the size of the army as they set out in 1812. Tufte and Graves-Morris point out that Minard's innovative graph relies on six variables: size (of army), latitude and longitude (where the army was), direction (that army was moving), location (at certain dates), and temperature (where the army was).[6]

One of the first effective means of using medical data to generate knowledge was developed in 1858 when Florence Nightingale used a polar-area diagram (also called a coxcomb chart) to demonstrate the relationship between sanitary conditions and soldiers' deaths compared to death from battlefield wounds (figure 2).[7 8]

Since then, standardized charts and graphs have been used for specific types of healthcare data to quickly determine the need for appropriate interventions. For example, graphing vital signs data can quickly identify a rise or fall in physiological data, indicating the need for an intervention and demonstrating the effectiveness of the intervention; and

To cite: West VL, Borland D, Hammond WE. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2014-002955

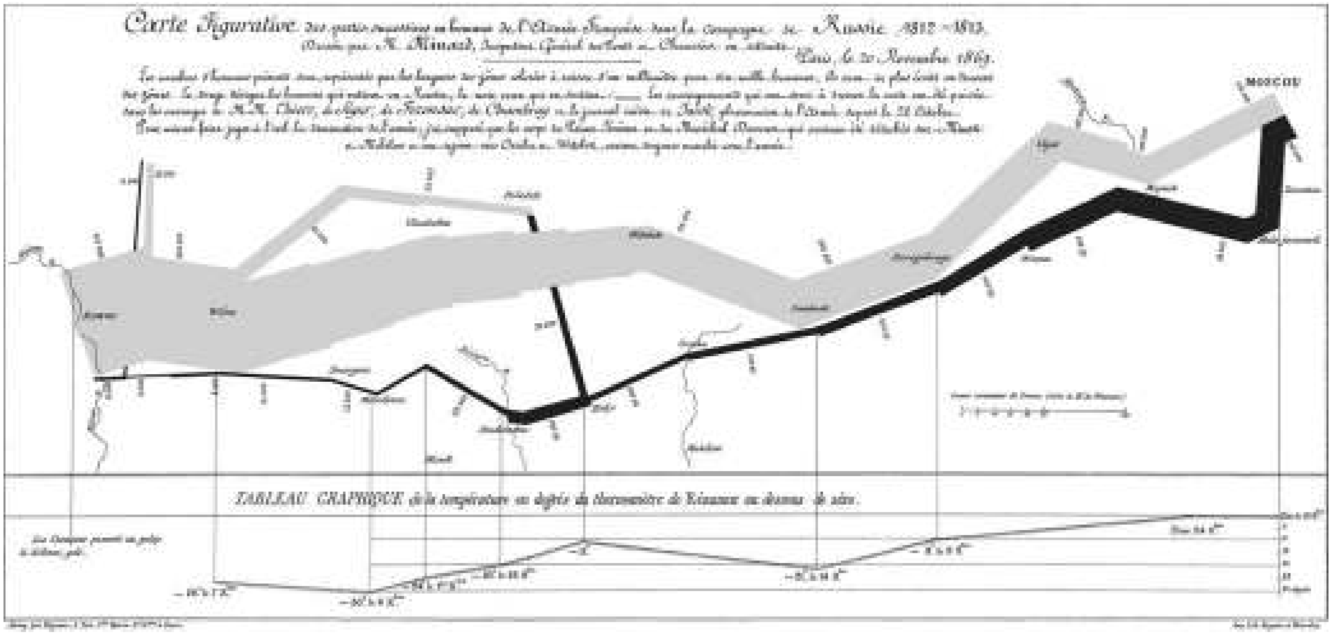


Figure 1 Translation: 'Figurative chart of the successive losses in men by the French army in the Russian campaign 1812–1813. Drawn up by Mr. Minard, inspector-general of bridges and roads (retired). Paris, 20 November 1869. The number of men present is symbolized by the broadness of the colored zones at a rate of 1 mm for ten thousand men; furthermore, those numbers are written across the zones. The red [note: gray band here] signifies the men who entered Russia, the black those who got out of it. The data used to draw up this chart were found in the works of Messrs. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished journal of Jacob, pharmacist of the French army since 28 October. To better represent the diminution of the army, I've pretended that the army corps of Prince Jerôme and of Marshall Davousz which were detached at Minsk and Mobilow and rejoined the main force at Orscha and Witebsk, had always marched together with the army.' Public domain (U.S.) image via Wikimedia Commons. Available at http://commons.wikimedia.org/wiki/File:Minard.png (accessed 21 July 2014).

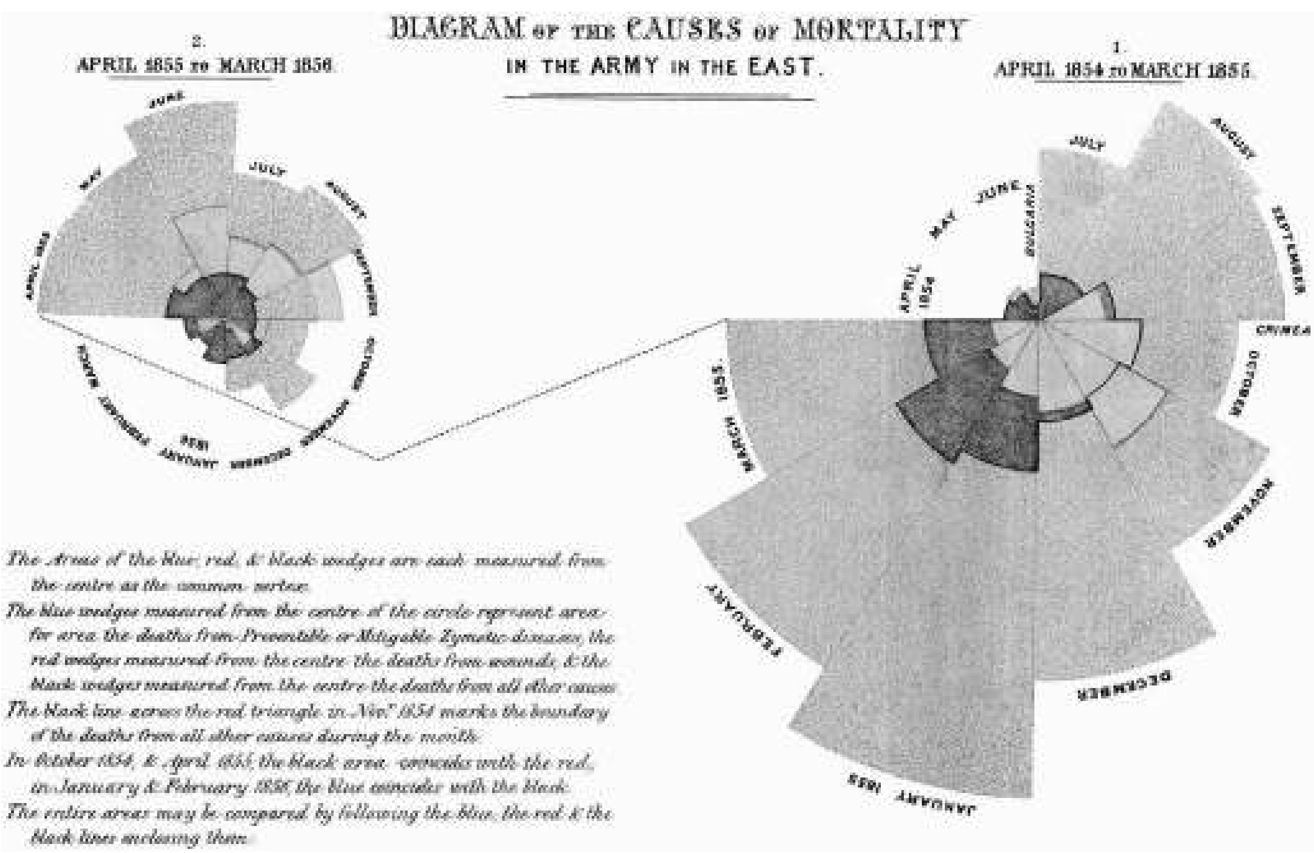


Figure 2 Florence Nightingale's coxcomb chart representing causes of death each month between April 1854 and March 1856 during the Crimean War. The *large outer gray bands* represent deaths attributed to lack of sanitation in the wards, the *lighter gray middle bands* to death from wounds during the war, and the *darkest inner bands* to other causes. Public domain (U.S.) image via Wikimedia Commons. Available at http://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg (accessed 21 July 2014).

Fishbone diagrams are commonly used graphic representations of laboratory results.

A plethora of scales, shapes, and colors have been used with both small and large datasets rendered as visual diagrams such as bar charts, line graphs, scatterplots, and pie charts to reveal patterns leading to knowledge discovery. Industries such as finance, accounting, and the petroleum industry routinely use information visualization, defined as 'interactive, visual representations of abstract data to amplify cognition',[9] using innovative approaches that account for both the volume and complexity of their data. In the healthcare field, however, applications of advanced visualization techniques to large and complex EHR datasets are limited.

Data in healthcare

In 1994, Powsner and Tufte[10] proposed summarizing patient status with test results and treatment data plotted on a graph. This was one of the earliest examples of using several diverse datasets in medical records to visualize information. Also in the 1990s, Plaisant *et al*[11] developed LifeLines as a means to visualize patient summaries using several different graphical attributes, for example colors and lines depicting a patient's discrete events. Furthermore, Shahar and Cheng[12 13] developed Knowledge-based Navigation of Abstractions for Visualization and Explanation (KNAVE) as a means to explore time-oriented, semantically-related concepts.

Clinical records by nature contain longitudinal data of patient visits over time, with records of changing problems, medications, treatments, and responses related to evolving health status. Graphs are routinely used to illustrate data in a way that comparisons, trends, and associations can be easily understood. In healthcare studies, the use of graphs with time as the horizontal axis to display various types of data has been increasing, and several well established visualization tools have been developed using temporal data, with LifeLines/LifeLines2[11 14–16] and KNAVE/KNAVE-II/VISITORS[17–20] the most widely reported. When querying 'longitudinal studies' in PubMed, 7071 publications were found in 1983, with the number consistently rising through the following 30 years to 45 821 studies published in 2013.

Longitudinal data from EHRs displayed through innovative visualization techniques has tremendous potential for discovering useful information in the data. Until health record data became widely available electronically, however, there was little emphasis on using such large and complex datasets. We argue that EHR data is actually a new kind of data that requires new visualization techniques beyond graphs and charts to accommodate the size of the dataset and explore the contents of the data.

Exploring EHR data with visualization techniques other than tables, graphs, and charts is a nascent approach to understanding the information in EHRs. A comprehensive monograph by Rind *et al*[21] focuses on a survey of visualization systems and criteria typically used by designers of systems. A book by Combi *et al*,[22] and two book chapters[23 24] also describe several of the visualization systems reported in the literature. We report our results from a systematic review that describes how innovative visualizations are being used with large and complex EHR data as a means to present or 'discover' information without specific hypotheses.

Objectives

The aim of this review is to investigate the visualization techniques that have been used with EHR data and answer the following questions:

- What is the prevalence of the use of information visualization with EHR data?
- Are techniques being used for knowledge discovery with an entire EHR dataset?
- What has been learned from research on visualization of EHR data?

METHODS

We conducted a systematic literature review following the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement.[25] Our review was limited to articles published between 1996 and 2013. We began with 1996, the year one of the largest healthcare systems in the USA, the Veterans Health Administration, first mandated the use of EHRs.[26] The Health Insurance Privacy and Portability Act (HIPPA) of 1996 was also enacted to provide security of individually identifiable health information, with a consensus that EHRs would be the most effective way to assure compliance with HIPPA. It is also the year when the first study using visualization with complex data (medical records histories and associated longitudinal data) was published by Plaisant *et al.*[11] This time interval enables us to construct the historical timeline for the use of information visualization in healthcare, particularly as data have become more common electronically due to the legislative requirement for conversion to EHRs.

An electronic literature search was conducted in May–July 2013 using MEDLINE, the most frequently used reference database in healthcare, and Web of Knowledge. This was supplemented using citation searching and gray literature searching. Reference lists from highly relevant articles were also reviewed to find additional articles. Broad keywords were used to assure a comprehensive document search (see table 1).

Inclusion and exclusion criteria

Articles had to include the use of EHR data using innovative visualization techniques, or describe developing techniques that would be applied to EHR data. We define EHR data as data in electronic clinical records that contain clinical information (eg, demographics, problems, treatments, procedures, medications, labs, images, providers) collected over time that can be shared among all authorized care providers. We define innovative visualizations as visualizations other than standard graphs traditionally used for displaying healthcare information (such as bar charts, pie charts, or line graphs) that use complex data, which we define as data with multiple types of variables and many data points, resulting in an exceptionally large amount of data, such as that in an entire EHR. We were interested in any innovative visualization technique for vast amounts of information that might be the foundation for an interactive system; therefore, although interaction is a key characteristic of information visualization, we included articles describing static visual

Table 1 Search terms used in search

Keyword	Boolean	Additional keywords
Information visualization		
Information visualization	AND	Health data, electronic health record, electronic medical record
Visualization	AND	Big data, clinical data, health data, health care data, healthcare data, electronic health record, electronic medical record

representations of large amounts of EHR data in addition to interactive visualizations.

Articles were excluded if they related to animals or plants, were position papers describing the need for visualizing data or ideas for techniques in visualization, or did not describe specific techniques used for the visualization or have figures showing the results from visualization. The literature is replete with articles on visualization in genetics, syndromic surveillance, and geospatial environmentally aware data, which we also did not include in our review because we were focused on clinical EHR data as defined above. There were many articles on the technical details related to visualization techniques, which did not fit our target for studies explaining how clinical data is used in visualization; these were also excluded.

Article selection and analysis

The authors, title, journal, year of publication, and abstract for each article were collected in an Excel spreadsheet. To identify key themes for matrix analysis, the first 50 abstracts and titles were reviewed; 11 themes were identified. These themes were then added to the spreadsheet to form a matrix for reviewing and categorizing all abstracts and to assist with determining which should be excluded.

After reviewing all abstracts and eliminating those categorized with exclusion criteria or lacking inclusion criteria, full articles of the remaining were read for eligibility. Our primary interest in conducting the study was to understand what innovative information visualization techniques in healthcare have been reported using EHR data since 1996. The review is not a meta-analysis and does not include a statistical analysis. The objective of the study was to investigate the prevalence of information visualization techniques used with EHR data, therefore we did not conduct a risk of bias assessment.

RESULTS

A total of 847 references were retrieved from our initial search of electronic databases, specifically MEDLINE (PubMed and PMC) and Web of Knowledge. A search of the gray literature and hand-searching references from articles yielded an additional 44 papers. All abstracts were reviewed, with duplicates removed (n=191). We then excluded 666 articles because the visualizations discussed were diagnostic, did not relate to EHR data, focused on animals or plants, used genomics data, discussed geospatial data or syndromic surveillance, were position papers suggesting the need for visualization or describing a potential visualization technique, or were primarily discussions of the technical details of visualization.

The full text of each of the remaining 34 articles was then read; 16 of these articles were excluded (table 2 lists for reasons for exclusion). Results of the screening process in the analysis are noted in the flow diagram in figure 3.

Eighteen articles were included in the qualitative synthesis. The online supplementary table S3 summarizes those included.[11 14–20 27–36]

The studies reviewed describe prototypes in various stages of development. Four of the articles describe LifeLines, the most advanced application, with its continued revisions and application in various populations. First described in 1996 by Plaisant *et al*,[11] LifeLines was developed as a prototype using data from the Maryland Department of Juvenile Justice to provide a visual overview of one juvenile's record on a single screen. LifeLines, using electronic health data, provides a timeline of a single patient's temporal events; time is represented on the horizontal axis, and events (problems, allergies, diagnoses, complaints, labs, imaging, medications, immunizations, communication) are listed vertically.

Table 2 Articles excluded from analysis

Reason for exclusion	No.	Explanation
Article not applicable	9	Articles are medical guidelines, no visualization is described, or articles describe process
Visualization not applicable	1	Does not use EHR data
Geospatial information	1	
Genetics	1	
Position paper	3	Ideas for visualization
Technical	1	
Total	16	

EHR, electronic health record.

LifeLines evolved to LifeLines2 and the use of multiple patient records. LifeLines2 research found that users want to see both numerical and categorical data, and that the ability to drill down into details when looking at patient records is an important feature.[16] Several other visualization techniques have been developed by this team using multiple records, for example LifeFlow,[31] developed for use with millions of patient records visualized on a single page that allows the user to see trends and evaluate quality of care. Using LifeFlow, new users can easily explore the data to understand patterns and trends at a high level.

Four articles[17–20] describe a second innovative visualization called VISITORS, or Visualization of Time-Oriented Records. VISITORS is based on earlier work of Shahar and colleagues, whose research conceptualized clinical data (eg, multiple measures of temperature over time) summarized into abstractions (in this case, fever). This was KNAVE[12]; KNAVE-II is a later enhancement.[18] Like LifeLines, VISITORS applies what researchers learned in earlier applications for a single record to develop an application that accommodates diverse temporal data from multiple records. Usability testing found the system feasible for exploring longitudinal data for quality or clinical results. The interface used with VISITORS was deemed to need simplification, in spite of the short time it took for users to learn how to use the system.

One article that we might have excluded used relatively simple linear graphs to illustrate the correlation of abstract concepts with laboratory values.[29] The data used in the analysis are from the 3 million patient EHRs for New York Presbyterian Hospital, promising complexity in the data. Both factors, EHR data and complex data, are inclusion criteria; therefore, the study was included in the final analysis. Seven laboratory tests and sign-out notes used primarily by residents to assist overnight staff caring for inpatients were abstracted. From the sign-out notes, 30 clinical concepts were identified using pattern matching, and then correlated with normalized lab values graphed on a timeline. The research showed the value of using time in the correlation, and the value of using aggregated data from many records versus a single record. It also demonstrates how temporal patterns can be visually found in EHR data using pattern matching and temporal interpolation.

A different approach is proposed by Joshi and Szolovits[34] using a radial starburst to show the complexity of data represented over a 100-dimensional space. The complexity is reduced by using machine learning to group similar clusters of patients

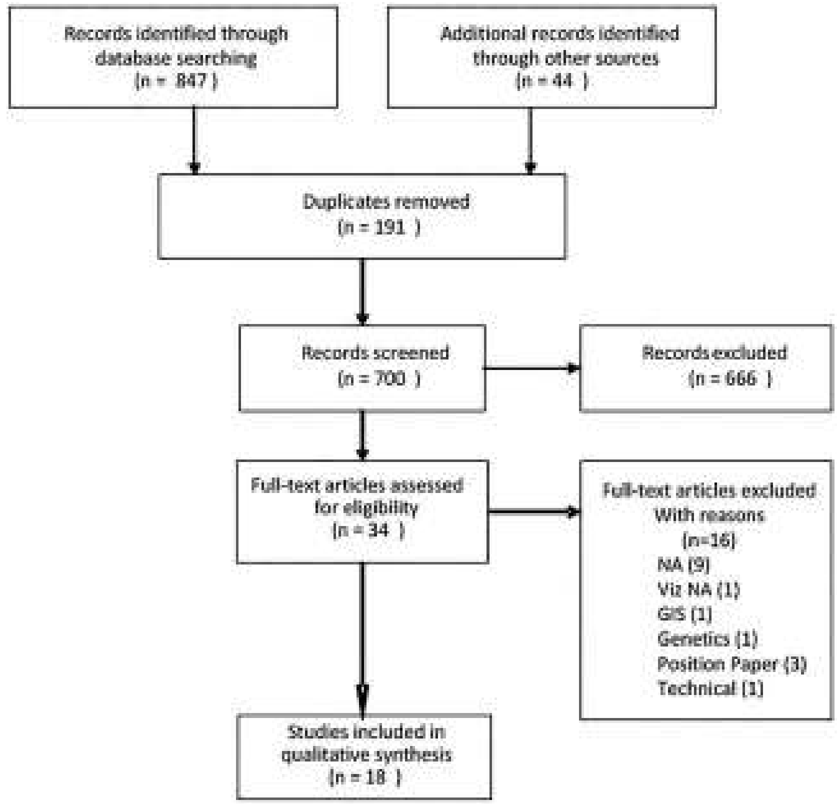


Figure 3 Flow of information through the different phases of systematic review. Adapted from the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) group.[25]

characterized by eight physiological foci to allow a user to look at one patient and evaluate the severity of that patient's condition. This is an example of using a very large set of data, or the EHR 'big data' as a clinical decision making tool. Although it is a static representation, Joshi and Szolovits provide a visual representation with an interesting presentation of complex data that has potential as a foundation for an interactive system with interactions such as filtering, selection, or brushing (highlighting a subset of data).

Gotz *et al*[30] developed Dynamic Icons, or DICON, as a visualization technique for exploring clusters of similar patients. By applying algorithms to EHR data, they found clusters of patients similar to the target patient. The user can interactively explore the clusters represented as icons on a treemap. They found this visual representation, a unique approach to visualization of healthcare data, required time for users to understand. Once users understood the design concept, however, the interface provided functionality for rapidly analyzing the data using icons that could be easily controlled.

Gotz and Wongsuphasawat[32 33] designed Outflow as a means to look at disease progression paths based on the assumption that the onset of a particular disease symptom applies perpetually, with common disease states among patients and transitions between the states. Outflow allows users to look at a visual display consisting of multiple events, their sequences, and outcomes to quickly analyze the event sequences in order and accurately identify factors most closely correlated with specific pathways.

Wang *et al* report using LifeLines2 and sentinel event data for subject recruitment to clinical trials.[19] They found using alignment, ranking, and filtering functions reduced user interaction time when working with sentinel events. Its use for subject recruitment was found to be questionable, however. Data in medical records can be somewhat uncertain, making the timeline inaccurate. For example, a patient with a long-standing diagnosis of asthma who visits a care provider for shortness of breath may be coded as first being diagnosed with asthma on that visit, even though the diagnosis of asthma was made previously. If a clinical trial includes patients diagnosed with asthma within a certain time range, the patient would be excluded in recruitment.

Fifteen studies address the use of temporal data.[11 14–20 27 29 31–35] Most articles describe interactive visualizations. All but two articles focus on use of the visualizations for clinical decision support. The two visualizations not used for decision support suggest use for quality assurance and improvement.[25 28]

Most studies that included an evaluation of the visualization described the training of the user and training time. One study reported training time of 6 min for its visualization, which used radial displays with a body map in the center of the radius and the relevant physiological parameters highlighted on the body map.[36] This was the shortest training time reported; the longest was a half hour.[30]

Although several ways to visualize EHR data are described, it was difficult to discern if the data as described were actually real-time data, or retrospective data or databases with predetermined datasets. Some of the articles describe systems for data visualization, for example, LifeLines2 and VISITORS. Others use visualization techniques such as sequential displays,[31 36] treemaps,[28 30] radial displays,[34 36 38] or icicle trees.[31]

DISCUSSION

Although most studies recognize the importance of the growing amount of clinical data, we found few innovative EHR visualization techniques that lend themselves to the large amount of data available electronically. Prior to 2010, seven publications we reviewed employed different and innovative visualization techniques with healthcare data; three of those describe LifeLines and three describe KNAVE-II/VISITORS. With the HITECH Act in 2009, national interest in EHRs was high, with increasing interest in knowledge that might be discovered by using visualization techniques applied to EHR data. Three studies on visualization of electronic health data were reported each year in 2010 and 2011, and four in 2012. Data from 2013 are not inclusive since our review was conducted in May–July 2013 (figure 4).

Several themes are common: the type of data accessible to the user, meaningfulness of visualizing large amounts of data, usability, and training time. Challenges from research to date can be broadly categorized into four areas identified by Keim *et al*[37] in other domains using very large, complex datasets: data (quality, size, diversity), users (needs, skills), design (intertwining both in a system that provides an easy way to visually explore and analyze results), and technology (tools, infrastructure).

Research on EHR visualization provides some important lessons on challenges that need to be addressed:

- The amount of EHR data and its display is a challenge; the more data, the more difficult it is to see and identify meaningful patterns in visualizations. Using tools such as zoom, pan, and filter reduces some of the clutter, but the purpose of the visualization will affect the use of such tools. If researchers are to use entire datasets from EHRs to discover information within the data, it will be necessary to develop better ways to manage the massive amounts of data.
- The size and complexity of EHR data is a challenge. Color, density, and filtering techniques are commonly used to distinguish variables or temporal events. Although scaling and zooming have been used to resize data, none of the reported techniques in the studies we reviewed discuss applicability to an entire EHR dataset and the potential for knowledge discovery in this very large composite dataset.
- The ability to use temporal data in visualizing aggregate data from EHRs is important to users.
- Researchers need to be cognizant of the many variables that can lead to uncertain data in EHRs; uncertain data can distort temporal events.
- EHR data are complicated by missing values, inaccurate data entry, and mixed data types that must be considered in developing visualization techniques.
- Presenting a great deal of information in a single screen shot where the user can interactively explore the information is an important design feature.
- Users want to see both categorical and numerical data when interactively exploring the data, and they like to look at the detail in the record. This is challenging with visualizing an extremely large amount of data in an EHR, but important for user satisfaction.
- A normalization scheme is needed for aggregated numerical data.
- The time it takes to learn the system is an important consideration that is complicated by the complexity of the data using visualizations that are different from those most clinicians and researchers are used to seeing, such as charts and graphs.
- Training time to understand and effectively use the visualization for its intended purpose should be considered when developing visualization techniques. Training is usually the user's first introduction to visualization. The complexity of the visualization and ways to navigate the display will increase training time if it is not easy to explain or demonstrate the functionality of the visualization.

Aigner *et al* have identified similar challenges working with temporal data, which is inherent in EHR data: the complexity, quality, diversity, and uncertainty of data; the interfaces and roles of the users; and evaluation of quality and effectiveness of the design.[38] The interest and challenges in data analysis with 'visual presentation and interaction technologies' that can be used with very large and complex datasets is universal.[39] The ability to explore and gain a deeper understanding of the value of 'big data' will encourage adoption of visualization techniques in healthcare. Research focused on these challenges is needed if we are to fully utilize EHR data for knowledge discovery.

Limitations

Although there are numerous articles published by Plaisant *et al* and Shahar and Klimov that are related to the techniques incorporated in their specific visualizations (LifeLines/LifeLines2/LifeFlow/EventFlow and KNAVE/KNAVE-II/ VISITORS), our review was limited to those articles that were the primary publications describing the innovative visualization technique and its application to electronic health data. By restricting our review to a narrow segment of this literature, we may have inadvertently eliminated meaningful details from our review.

Our search terms were intentionally broad; we eliminated articles whose abstracts indicated the articles were more technical in nature, and we eliminated articles whose focus was on geospatial representation. We may have obtained different results had more specific terms been used.

Finally, there are books and book chapters that deal with visualization of healthcare data. These types of publications are not

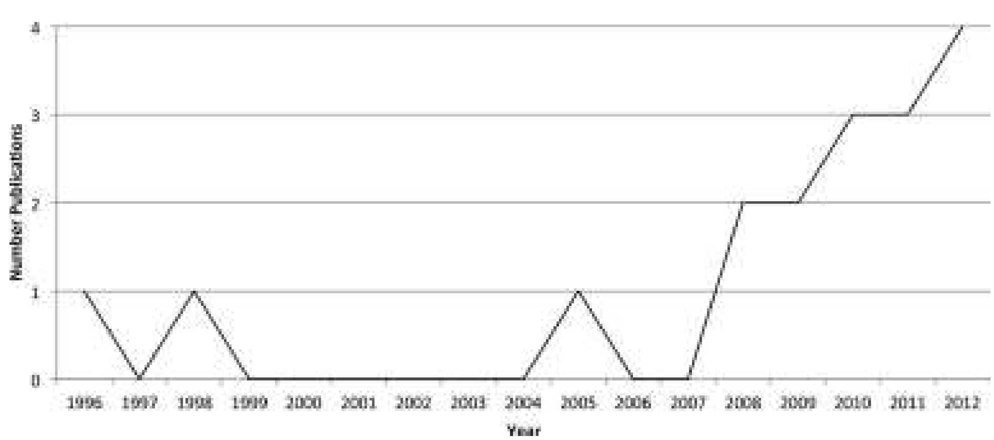


Figure 4 Number of publications included in review.

included in our review, but may contain information relevant to this review.

CONCLUSIONS

This study was conducted to determine the prevalence of the use of information visualization for EHR data, what techniques have been used, and what research has taught us to date. Although there is increasing interest in visualization of electronic healthcare data, few techniques have been found to effectively and efficiently display the large and complex data in EHRs.

The new buzzword in healthcare is 'big data', often used in conjunction with data analysis. Most studies have found that visualization of EHR data requires techniques that will handle not only 'big data', but the temporal complexity of constantly changing variables found within EHR data. Disciplines such as computer science, engineering, and genetics have developed visualizations to improve presentation, analysis, and understanding of data. The healthcare provider community has not yet taken advantage of these methods or significantly explored the use of new visualization techniques to accelerate the use and understanding of EHR data. We have identified important findings reported in the literature that can help guide future research needed to explore, refine, and retest visualization techniques. Only then will stakeholders begin to take advantage of the wealth of knowledge within EHR data.

Acknowledgements The authors wish to acknowledge Rene' Hart, Staff Assistant at the Duke Center for Health Informatics, for her help in organizing the literature review.

Contributors VLW conducted the literature search. All authors contributed to the analysis and text. VLW and DB provided the figures. All edited, reviewed, and approved the final version.

Funding This work was supported by the US Army Medical Research and Materiel Command (USAMRMC) grant number W81XWH-13-1-0061.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Bush GW; Office of the Press Secretary, the White House. Executive Order: Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. Press release, April 27, 2004. http://www.whitehouse.gov/news/releases/2004/04/print/20040427-4.html2004 (accessed 23 Jul 2013).
2 Bush GW. State of the Union Address, Promoting Innovation and Competitiveness, President Bush's Technology Agenda. 2004.
3 Data Show Electronic Health Records Empower Patients and Equip Doctors. Press release, July 17, 2013. http://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-Releases/2013-Press-Releases-Items/2013-07-17.html (accessed 23 Jul 2013).
4 Spence I. William Playfair and the psychology of graphs. *American Statistical Association JSM Proceedings*; 2006:2426–36.
5 Playfair W. The statistical breviary; shewing, on a principle entirely new, the resources of every state and kingdom in Europe; illustrated with stained copper plate charts, representing the physical powers of each distinct nation with ease and perspicuity. To which is added, a similar exhibition of the ruling powers of Hindoostan. London: J Wallis, 1801.
6 Tufte ER, Graves-Morris PR. *The visual display of quantitative information.* Vol 2. Cheshire, CT: Graphics Press, 1983.
7 Nightingale F. Notes on matters affecting the health, efficiency, and hospital administration of the British Army. Founded chiefly on the experience of the late War. Presented by request to the secretary of state for War. Privately printed for Miss Nightingale, Harrison and Sons, 1858.
8 Lienharg J. The Engines of Our Ingenuity, Episode 1712: Nightingale's Graph. Podcasts between 1988–2002. 1988–2002. http://www.uh.edu/engines/epi1712.htm (accessed 21 Jul 2013).
9 Card SK, Mackinlay JD, Shneiderman B, eds. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999.
10 Powsner S, Tufte E. Graphical summary of patient status. *Lancet* 1994;344:386–98.
11 Plaisant C, Milash B, Rose A, *et al*. LifeLines: visualizing personal histories. *SIGCHI Conference on Human Factors in Computing Systems Proceedings*; 1996:221–7.
12 Shahar Y, Cheng C. Intelligent visualization and exploration of time-oriented clinical data. Systems Sciences, HICSS-32, 32nd Annual Hawaii International Conference Proceedings 1999 (Volume:Track4).
13 Shahar Y, Cheng C. Model-based visualization of temporal abstractions. *Comput Intell* 2000;16:279–306.
14 Plaisant C, Mushlin R, Snyder A, *et al*. LifeLines: using visualization to enhance navigation and analysis of patient records. *AMIA Symposium Proceedings*; 1998:76–80.
15 Wang TD, Plaisant C, Quinn AJ, *et al*. Aligning temporal data by sentinel events: discovering patterns in electronic health records. *CHI '08 SIGCHI Conference on Human Factors in Computing Systems Proceedings* 2008:457–66.
16 Wang TD, Wongsuphasawat K, Plaisant C, *et al*. Visual information seeking in multiple electronic health records: design recommendations and a process model. *1st ACM International Health Informatics Symposium Proceedings*; 2010:46–55.
17 Klimov D, Shahar Y. A framework for intelligent visualization of multiple time-oriented medical records. *AMIA Annu Symp Proc* 2005;2005:405–9.
18 Martins SB, Shahar Y, Goren-Bar D, *et al*. Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data. *Artif Intell Med* 2008;43:17–34.
19 Klimov D, Shahar Y, Taieb-Maimon M. Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records. *Methods Inf Med* 2009;48:254–62.
20 Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artif Intell Med* 2010;49:11–31.
21 Rind A, Wang TD, Aigner W, *et al*. Interactive information visualization to explore and query electronic health records: a systematic review. *Foundations Trends Hum-Comput Interact* 2013;5:207–98.
22 Combi C, Keravnou-Papailiou E, Shahar Y. Temporal information systems in medicine. Springer, 2010.
23 Aigner W, Kaiser K, Miksch S. Visualization techniques to support authoring, execution, and maintenance of clinical guidelines. Computer-based Medical Guidelines and Protocols: A Primer and Current Trends 2008;139:140–59.
24 Lesselroth BJ, Pieczkiewicz DS. Data visualization strategies for the electronic health record. Nova Science Publishers, Inc., 2011.
25 Moher D, Liberati A, Tetzlaff J, *et al*.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
26 Anderson M. Lessons learned from the Veterans Health Administration. Healthc Pap, 5, 30–37. 2005. https://www.thecsiac.com/sites/default/files/files/Clinger%20Cohen%20(1996).pdf (accessed 15 Apr 2013).
27 Bashyam V, Hsu W, Watt E, *et al*. Problem-centric organization and visualization of patient imaging and clinical data. *Radiographics* 2009;29:331–43.
28 Willison B. Advancing Meaningful Use: Simplifying Complex Clinical Metrics Through Visual Representation. Parsons Institute for Information Mapping (PIIM) Research 2010.
29 Hripcsak G, Albers DJ, Perotte A. Exploiting time in electronic health record correlations. *J Am Med Inform Assoc* 2011;18(Suppl 1):i109–115.
30 Gotz D, Sun J, Cao N, *et al*. Visual cluster analysis in support of clinical decision intelligence. *AMIA Annu Symp Proc* 2011;2011:481–90.
31 Wongsuphasawat K, Guerra Gómez JA, Plaisant C, *et al*. LifeFlow: visualizing an overview of event sequences. *SIGCHI Conference on Human Factors in Computing Systems Proceedings* 2011:1747–56.
32 Gotz D, Wongsuphasawat K. Interactive intervention analysis. *AMIA Annual Symposium Proceedings*; 2011:274.
33 Wongsuphasawat K, Gotz D. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Trans Vis Comput Graph* 2012;18:2659–68.
34 Joshi R, Szolovits P. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. *AMIA Annu Symp Proc* 2012;2012:1276–1283.
35 Stubbs B, Kale DC, Das A. Sim•TwentyFive: an interactive visualization system for data-driven decision support. *AMIA Annu Symp Proc* 2012;2012:981–900.
36 Zhang Z, Wang B, Ahmed F, *et al*. The five W's for information visualization with application to healthcare informatics. *IEEE Trans Vis Comput Graph* 2013;19:1895–1910.
37 Keim DA, Kohlhammer J, Ellis G, *et al*., eds. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010.
38 Aigner W, Federico P, Gschwandtner T, *et al*. Challenges of Time-oriented Data in Visual Analytics for Healthcare. *IEEE VisWeek Workshop on Visual Analytics in Healthcare*; 2012.
39 Thomas JJ, Cook KA. A visual analytics agenda. *IEEE Comput Graph Appl* 2006;26:10–3.

Innovative information visualization of electronic health record data: a systematic review

Vivian L West, David Borland and W Ed Hammond

J Am Med Inform Assoc published online October 21, 2014
doi: 10.1136/amiajnl-2014-002955

Table 3. Summary of articles included in qualitative systematic review.

No.	Authors	Title/Journal	Year	Proto type	Data	Population Developed For	Technique	Comments
1	Plaisant C, Milash B, Rose A, Widoff S, Shneiderman B[11]	LifeLines: visualizing personal histories/ Proc, CHI'96	1996	Yes	Single medical record, used sample data	Envisioned for all types of patients	One of the earliest reports on using graphical timelines to visualize multiple data in medical records.	Encoding decisions were a challenge. There was early recognition of the importance of access to accurate records. Established the ability to visualize a variety of data.
2	Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman, B[14]	LifeLines: using visualization to enhance navigation and analysis of patient records/ Proc AMIA Symp	1998	Yes	Single patient record, actual data	All patients	Reports on the status of LifeLines, described in Plaisant, et. al. (No. 1 above)	Longitudinal data increases the complexity of the data, making the vertical display of the timeline more difficult. Visualizing data in one display allows the user to see correlations and relationships among the variables.
3	Klimov D, Shahar Y[17]	A framework for intelligent visualization of multiple time-oriented medical records/ Proc AMIA Symp	2005	Yes	Multiple patient records	Patients with chronic diseases	Developed *Visualization of Time-Oriented Records* (VISITORS) using temporal abstraction from a large number of patient records, an enhanced version of KNAVE-II (KNAVE developed in late 1990s). Output is visualized as chart graphs and bands; supports both calendar and events timelines. Interactive.	Describes the VISITORS architecture. Objectives for VISITORS: aggregation of multiple patients over time for clinical research, better care for chronically ill patients, and QA. Includes both individual patient and population of patients, their raw data, and aggregated values that can be manipulated temporally (such as mean/min/max of lab values over hour/day/week). To implement, need domain-specific knowledge, plan to test using clinical databases.

No.	Authors	Title/ Journal	Year	Proto type	Data	Population Developed For	Technique	Comments
4	Martins SB, Shahar Y, Goren-Bar D, Galperin M, Kaizer H, Basso LV, McNaughton D, Goldstein MK[18]	Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data/ Art Intel in Med	2008	Yes	Single or small groups of patients using longitudinal data from many sources	Oncology patients	Single or small groups of patients using longitudinal data from many sources	KNAVE-II successfully tested simple concepts and features. Queries of data much faster and more accurate than when working with paper charts of spreadsheets. User can explore raw data or concepts represented by the data. Preparing for larger clinical studies with more complex data.
5	Wang T, Plaisant C, Quinn A, Stanchak R, Shneiderman B, Murphy S[15]	Aligning temporal data by sentinel events: discovering patterns in electronic health records/ SIGCHI Conf on Human Factors in Computing Systems	2008	Yes	Multiple patient records	Patient recruitment for clinical trials	Uses LifeLines 2 to incorporate sentinel event data (categorical data) as timelines, important for clinical trial subject recruitment according to inclusion/exclusion criteria, and for observational research. Incorporates align, rank, and filter interactions.	Visualizations aligning sentinel events in large complex datasets significantly decreases time to review and identify patterns of interest, and supports temporal comparisons of clinical events. Interface is quick to learn. Because of uncertainty in clinical data, sentinel events may not lend themselves to meaningful visual representations.
6	Klimov E, Taieb-Maimon M[20]	Intelligent interactive visual exploration of temporal associations among multiple time-oriented patient records/ Methods Inf Med	2009	Yes	Multiple patient records	All types of patients	Describes development of Temporal Association Charts (TAC) to display raw data and a domain concept, e.g. hemoglobin with time stamps. Allows user to interactively explore concepts and clinical data over time.	An enhancement used with VISITORS (Klimov and Shahar, No. 3 above). TACS were found to be functional in usability testing. Use of TACS relies on users to determine what to explore, so goals must be pre-determined versus made obvious through the visualization.

No.	Authors	Title/ Journal	Year	Proto type	Data	Population Developed For	Technique	Comments
7	Bashyam V, Hsu W, Watt E, Bui AA, Kangarloo H Taira RK[27]	Problem-centric organization and visualization of patient imaging and clinical data/ Radiographics	2009	Yes	Entire medical records	Radiology, neuro-oncology; authors suggest it can be applied to other clinical areas	Uses natural language processing to explore data for clinically relevant data that is linked to the anatomy of interest, which the user then explores.	The visualization itself is not unique, the ability to see textual data linked to the anatomic visualization is unique. Data is parsed from records using filters that provide relevant temporal, spatial, existential, and causal data. Usability testing is needed.
8	Willison B[28]	Advancing meaningful use: simplifying complex clinical metrics through visual representation/ Parsons Institute for Info Map (PIIM) Research	2010	Yes	Multiple patient records	Any type of patients	Uses treemaps and radial graphs to explore clinical data and compare to data in patient registries.	Examples show that visualization methods other than traditional methods improves understanding of the data and quickly identify areas of concern. Also suggests that effective visualization techniques differ based on the type of user and data (e.g. registry or QA).
9	Klimov D, Shahar Y, Taieb-Maimon M[19]	Intelligent visualization and exploration of time-oriented data of multiple patients/ Artif Intell Med	2010	Yes	Multiple patient records	Any types of patients	Paper is a more in-depth discussion of 2009 article by Klimov, et. al. (No. 6 above), focusing on visualization itself, interactivity, and temporal analysis.	User evaluation determined that the interface was complex at first, but less so after training and some exploration of the system. Improvements might be eliminating some of the features not used often, and including a visualization for comparison of population groups.

No.	Authors	Title/ Journal	Year	Proto type	Data	Population Developed For	Technique	Comments
10	Wang T, Wongsuphasawat K, Plaisant C, Shneiderman B [16]	Visual information seeking in multiple electronic health/ Proc of 1st ACM Internat Health Info Sympvf	2010	ContInuation	Multiple patient records	Any types of patients	Eight case studies using LifeLines2. See previous Wang et al. (No. 5 above)	Eight case studies over 2 1/2 years revealed strengths and weakness of LifeLines 2 for temporal categorical data. Studies show system led to discoveries that helped improve care. Conclusions acknowledge work is needed to move information visualization techniques forward. Six recommendations are provided for development of visualization tools in the future.
11	Hripcsak, G, Albers DJ and Perotte A[29]	Exploiting time in electronic health record correlations/ J Am Med Inform Assoc	2011	Yes	Population of patients, specifically data warehouse	Any types of patients	Temporal interpolation used with clinical concepts, e.g. diseases or symptoms, correlated with actual lab values to demonstrate the usefulness of looking at data and clinical processes together.	Shows that temporal properties in data with clinical associations can be easily extracted from EHR data. The most reliable information came from exploiting time. Using patients as their own controls to normalize the data seemed to reduce bias.
12	Gotz D, Sun J, Cao N, Ebadollahi S[30]	Visual cluster analysis in support of clinical decision intelligence/ AMIA Annu Symp Proc	2011	Yes	Population of patients	EHR data for many patients as a group	Uses *Dynamic Icons* (DICON), a visualization tool using treemaps to display clusters of similar patients representing multiple data points. Cluster analysis algorithms are used with an EHR to identify clusters. Each cluster is viewed interactively.	Tool being developed for clinical decision support. Can explore multidimensional data within icons. Does not require extensive training, and interface is intuitive. The embedded detail within the icons increases complexity. Refinements are needed.

No.	Authors	Title/ Journal	Year	Proto type	Data	Population Developed For	Technique	Comments
13	Wongsuphasawat K, Guerra Gómez JA, Plaisant C, Wang TD, Taieb-Maimon M, Shneiderman B [31]	LifeFlow: visualizing an overview of event sequences/ Proc of SIGCHI Confer on Human Factors in Comput Systems	2011	Yes	Multiple patients	ER patients	Using a hierarchical structure, records are aggregated according to sequential events to visualize prevalence and time between events.	Is scalable, can accommodate millions of records on one screen. Shows the user outliers in event data. Easy to calculate averages in temporal events. Can visualize two datasets on one screen to compare data. Overview of data allows for rapid inspection of data, potentially useful for QA
14	Gotz D, Wongsuphasawat K[32]	Interactive intervention analysis/ AMIA Annu Symp Proc	2012	Yes	EMR database	Similarity-based patient cohorts	Text analysis of unstructured data to identify similarity-based cohort; data converted to structured data using annotators stored with structured data. Data is visualized as disease-progression graphs	Allows user to see disease progression, interventions, and associated outcomes. Conclusion states the need for additional work. Further described in Wongsuphasawat and Gotz below. [No. 15}
15	Wongsuphasawat K, Gotz D [33]	Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization/ IEEE Vis and Comput Graphics	2012	Yes	Multiple patients	Cohorts of patients	Outflow visualization technique aligns temporal data from multiple patients by events, with the ability to explore the data along pathways, evaluate the sequence of events over time and outcomes. Information is visualized through nodes, layers, color coding, and edges between the transitions along pathways. See Gotz and Wongsuphasewat [No. 14] for additional information.	Users found visualization simple and the ability to see various outcomes along alternate pathways powerful. Large time gaps can make it difficult to follow the pathways. Events that can effect outcomes, such as medication to control triggers for specific outcomes, are not incorporated in the system yet. Design limitations are identified that need refinement.

No.	Authors	Title/ Journal	Year	Proto type	Data	Population Developed For	Technique	Comments
16	Joshi R, Szolovits P[34]	Prognostic physiology: modeling patient severity in intensive care units using radial domain folding/ AMIA Annu Symp Proc	2012	Yes	Single and multiple patients, ICU data	ICU patients	Uses a starburst to display clinical values related to one patient and Radial Domain Folding (RDF) for clusters of patients by organ state.	Suggested framework has application to large data sets such as EHRs, may be applicable to a variety of clinical settings. Visual representations have the potential to be the foundation for an interactive visualization system.
17	Stubbs B, Kale DC, Das A[35]	Sim•TwentyFive: an interactive visualization system for data-driven decision support/ AMIA Annu Symp Proc	2012	Yes	EHR data	Pediatric ICU patients	Uses multiple clinical data points to model patient similarities for decision analysis. Visualized using similarity plots, details-on-demand tables, and timeline charts.	Web-based system for clinical decision support, treatment planning, and physician education. Animation frame rate and loading graphics tested using several platforms (3) and browsers (4) with variable successes. System appears robust; feedback from testing indicates work is needed on usability. Has limited display of 25 patients and is not scalable. Limited HIPAA data display without secure server.
18	Zhang Z, Wang B, Ahmed F, Ramakrishnan I, Zhao R, Viccellio A, Mueller K[36]	The five W's for information visualization with application to healthcare informatics/ IEEE Trans Vis Comput Graph	2013	Yes	Single patient, eventually multiple patient data	A single patient, eventually comparison with a cohort.	A body map in the middle of a radial display and a multi-stage flow chart are used to drill down on data from a single patient record. Nodes in the radial display provide their hierarchical relationship and display additional data.	Filtering techniques are needed for large data sets: they are difficult to scale and browse. Cohort identification for clinical decision support would be a plus. Data uncertainty is shown using node saturation. Connections to online databases, e.g. drug information, support users.

Appendix I

Innovative Information Visualization of Electronic Health Record Data: a Systematic Review

Vivian West, David Borland, W. Ed Hammond

February 5, 2015

Duke Center for Health Informatics

Outline

- Background
- Objective
- Methods & Criteria
- Analysis & Findings
- Limitations
- Conclusions

West, V.L., D. Borland, and W.E. Hammond, Innovative information visualization of electronic health record data: a systematic review. JAMIA, Published Online First: 2014.

Background

- 2004 Presidential executive order: ‘Electronic Health Records (EHR) for all Americans
  - Provide accessible EHR for most Americans within 10 years
- 2009 Health Information Technology for Economic and Clinical Health Act (HITECH Act)
  - Allocated $19.2 billion in incentives to increase use of EHRs
- Centers for Medicare and Medicaid Services (CMS) incentive payments from CMS for adopting EHRs

eHR INCENTIVE PROGRAM

[1] Center for Medicare & Medicaid Services. [Electronic Health Records Incentive Programs Logo] Retrieved from http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms/

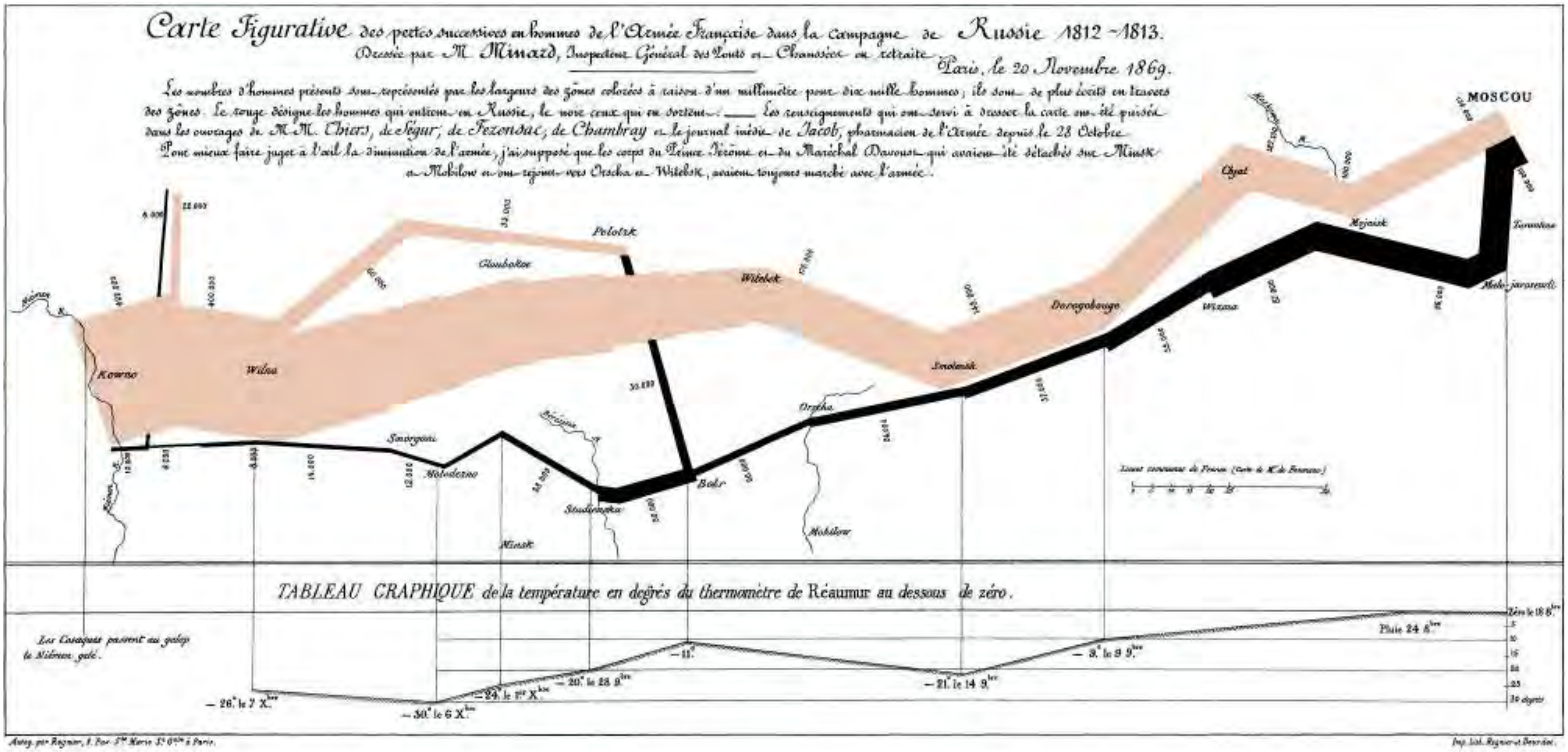
History of Data Visualization

- William Playfair
  - First to use basic graphical visualization of data
  - Stated readers *best understand and retain information by graphical representations of data*
  - *1786: The statistical breviary: shewing, on a principle entirely new, the resources of every state and kingdom in Europe; illustrated with stained copper plate charts, representing the physical powers of each distinct nation with ease and perspicuity*

[2] *Calton Hill - William Playfair*. [Digital image]. *Edinburgh World Heritage*. Edinburgh World Heritage, n.d. Web. 28 Jan. 2015. Retrieved from <http://www.ewht.org.uk/news/188/160/Calton-Hill---William-Playfair>.

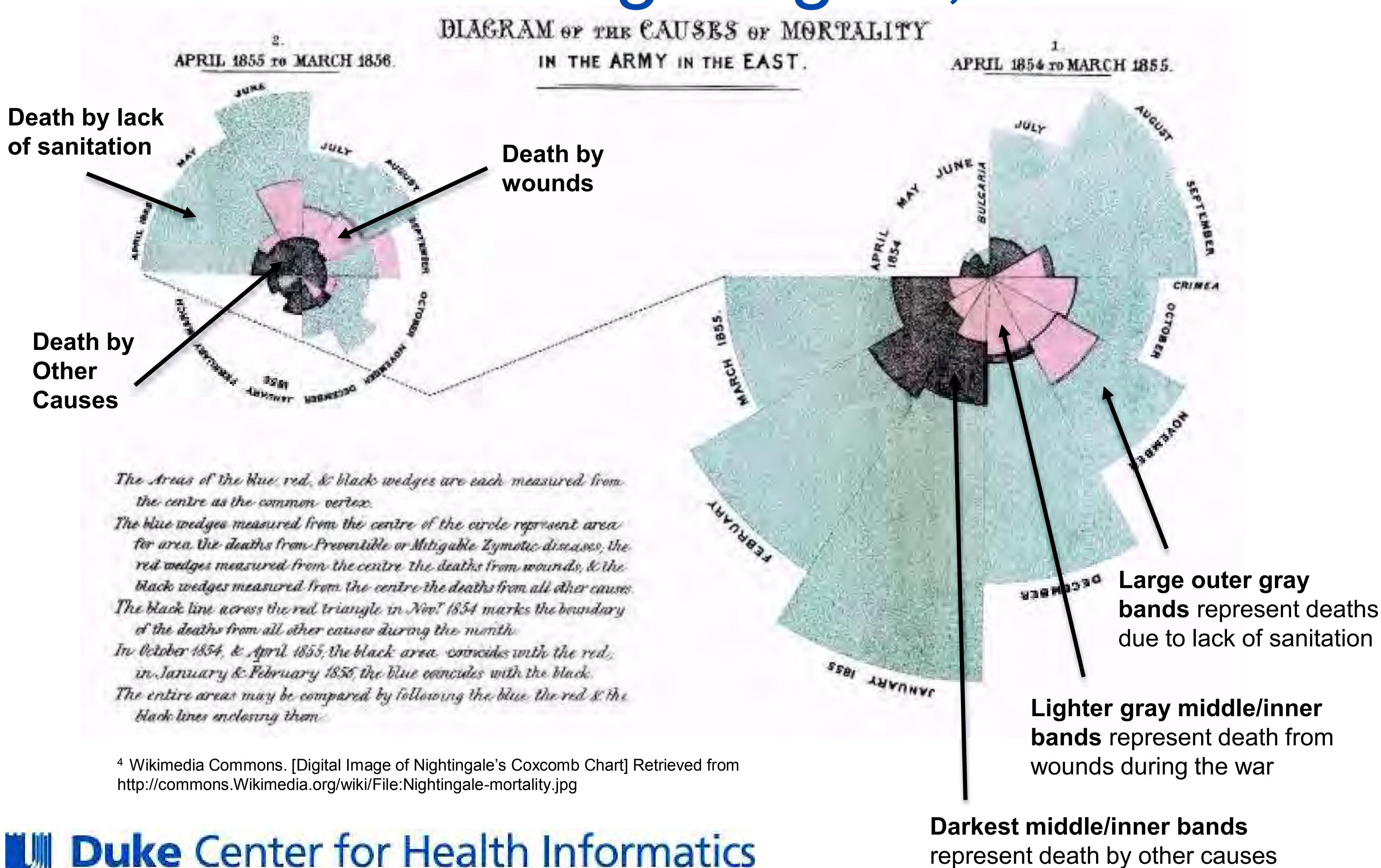
Napoleon’s March on Russia in 1812

Published by Charles Minard in 1861



[3] Wikimedia Commons. [Digital Image of Napoleon’s March on Russia] Retrieved from http://commons.wikimedia.org/wiki/File:Minard.png.

Florence Nightingale, 1858



4 Wikimedia Commons. [Digital Image of Nightingale's Coxcomb Chart] Retrieved from http://commons.Wikimedia.org/wiki/File:Nightingale-mortality.jpg

Duke Center for Health Informatics

Visualization Techniques

- Graphs for vital signs
- Fishbone diagrams: laboratory results
- Knowledge discovery using
  - Large and small data sets
  - Scales, shapes, colors
  - Bar chars, line graphs, scatterrplots, pie charts
- “Information visualization”-- Interactive visual representations of abstract data to amplify cognition
- Finance, accounting, and the petroleum industry: account for volume and complexity of data
- Visualization techniques to large and complex EHR datasets limited

Health Care Data Visualizations in the 1990's

- Summarize patient status
- Use several diverse data sets in EHR to visualize information
- Plots of test results & treatment data (Pownser & Tufte, 1994)
- Clinical records contain longitudinal data of patient visits over time with records of changing problems, medications, treatments, and responses related to health status
- Graphs illustrate data so comparisons, trends, and associations can be understood
- Healthcare studies use graphs with time as the horizontal axis
  - Visualization tools developed using temporal data

LifeLines(Plaisant et. al.)[4]

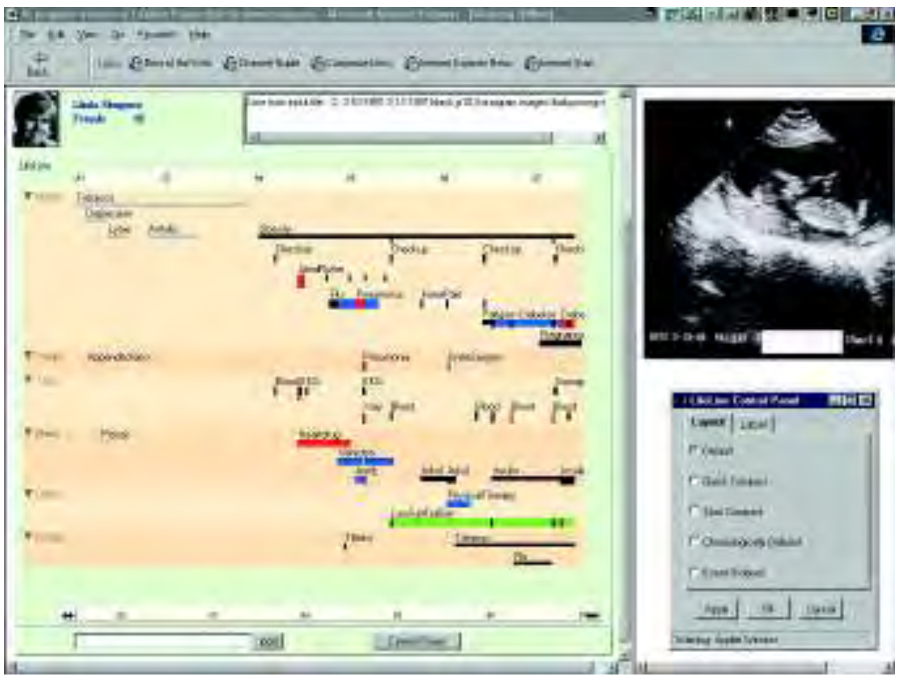
- Began with user interfaces for Juvenile Justice Information Systems
- Graphical attributes
  - Colors and lines depicting a patient's discrete events

[5] Lifelines (Original): User Interfaces for Juvenile Justice Information Systems. [Digital image]. *Human-Computer Interaction Lab*. University of Maryland Institute for Advanced Computer Studies, 2014. Web. 28 Jan. 2015. Retrieved from: <http://www.cs.umd.edu/hcil/temporalviz//>.

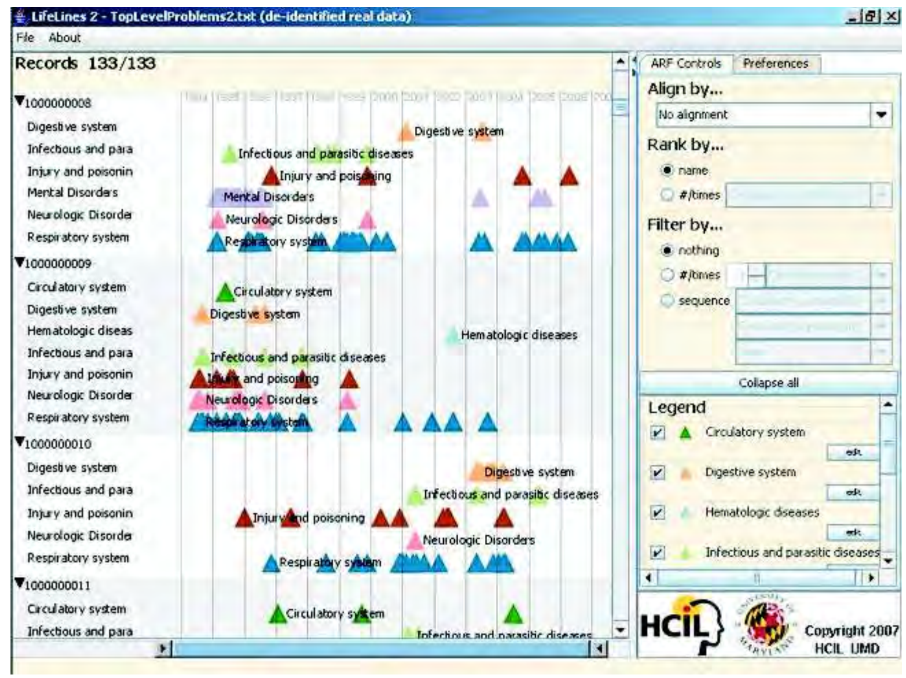
LifeLines/LifeLines2

Plaisant, et. al.

Visualizing Patient Records



Discovering Temporal Categorical Patterns Across Multiple Records



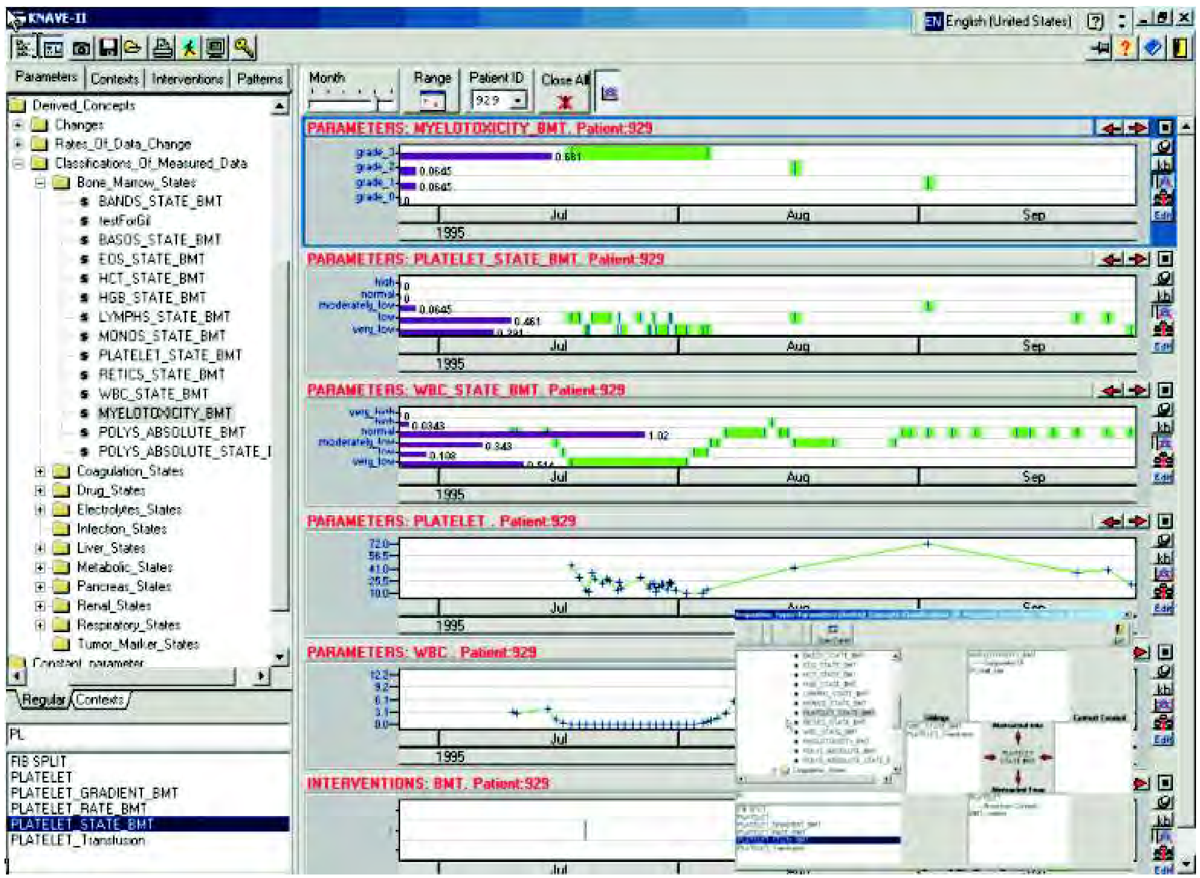
[6] LifeLines: Visualizing Patient Records. [Digital image]. *Human-Computer Interaction Lab*. University of Maryland Institute for Advanced Computer Studies, 2014. Web. 28 Jan. 2015. Retrieved from: <http://www.cs.umd.edu/hcil/temporalviz//>.

[7] Lifelines2: Discovering Temporal Categorical Patterns Across Multiple Records. [Digital image]. *Human-Computer Interaction Lab*. University of Maryland Institute for Advanced Computer Studies, 2014. Web. 28 Jan. 2015. Retrieved from: <http://www.cs.umd.edu/hcil/temporalviz//>.

KNAVE/KNAVE-II

Shahar & Cheng, Shahar, et. al.

Knowledge-based **N**avigation of **A**bstractions for **V**isualization and **E**xplanation



[8] *Research Project - KNAVE II*. [Digital image]. *Research Project - KNAVE II*. Ben Gurion University, n.d. Web. 28 Jan. 2015. <http://medinfo.ise.bgu.ac.il/medLab/ResearchProjects/RP_KNAVE.htm>.

Visualizing Health Care Data

- Longitudinal data from EHRs displayed through innovative visualization techniques has tremendous potential for discovering useful information in data
- Before EHR, little emphasis on using large and complex datasets
- LifeLines/LifeLines2 & KNAVE/KNAVE-II/VISITORS (**Vis**ualization of **T**ime-**O**riented **R**ecords) most widely reported
- Longitudinal studies in PubMed have increased from 7,071 publications (1983) to 45,821 (2013)
- EHR data is a new kind of data the requires new visualization techniques to discover knowledge

Objectives

- Investigate visualization techniques that have been used with EHR data and answer the following questions:
  - What is the prevalence of the use of information visualization with EHR data?
  - Are techniques being used for knowledge discovery with an entire EHR dataset?
  - What has been learned from research on visualization of EHR data?

Methods

- Conducted a systematic literature review following PRISMA (Preferred Reporting Items for Systematic reviews and Meta-analysis)
- Limited articles between 1996 and 2013
  - Veterans Health Administration first mandated use of EHRs
  - Health Insurance Privacy and Portability Act (HIPPA) enacted
  - First study using visualization with complex data published by Plaisant, et. al.

Methods

- Electronic literature search conducted in May – July 2013 using MEDLINE & Web of Knowledge
- Supplemented by citation searching and gray literature searching
- Used broad keywords to assure comprehensive document search

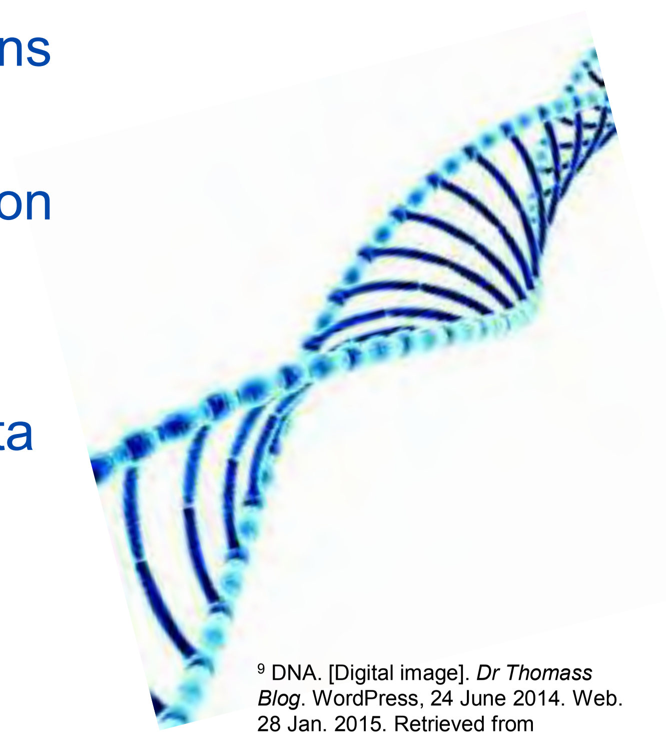
Keyword	Boolean	Additional keywords
Information visualization		
Information visualization	AND	Health data, electronic health record, electronic medical record
Visualization	AND	Big data, clinical data, health data, health care data, healthcare data, electronic health record, electronic medical record

Inclusion Criteria

- Use EHR data and innovative visualization techniques or describe techniques applicable to EHR data
- Includes articles describing static and interactive visualizations of EHR data
- Definitions:
  – EHR data – electronic clinical records containing clinical information (eg, demographics, problems treatments, procedures, medications, labs, images, providers)
  – Innovative visualizations – visualizations other than standard graphs; must use complex and large data

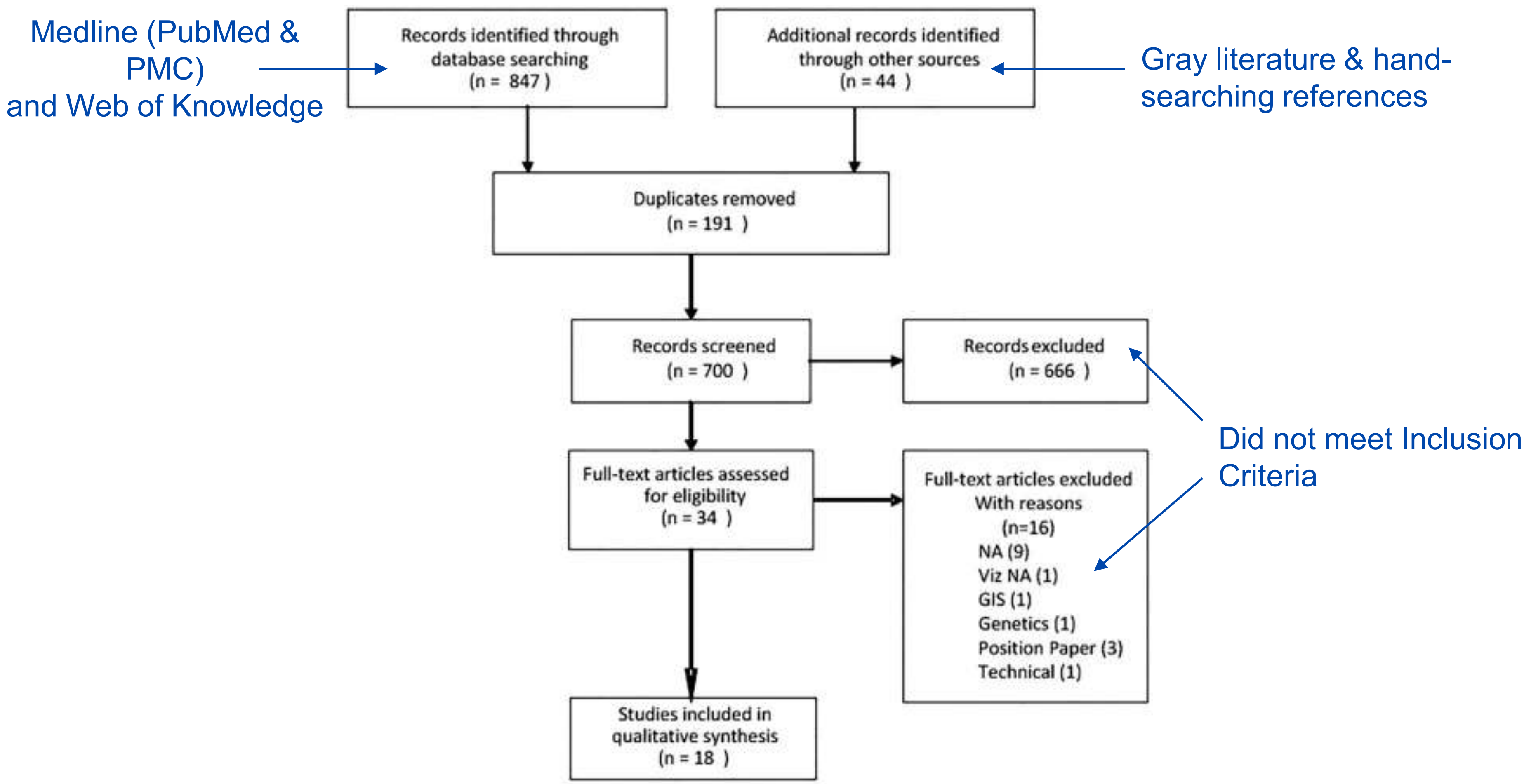
Exclusion Criteria

- Articles are excluded if:
  - Related to animals or plants
  - Did not describe specific techniques for visualization
  - Solely described the need for visualizations
  - Focused on non-EHR topics:
    - Technical details related to visualization
    - Genetics
    - Syndromic surveillance
    - Geospatial environmentally aware data



[9] DNA. [Digital image]. *Dr Thomass Blog*. WordPress, 24 June 2014. Web. 28 Jan. 2015. Retrieved from <http://www.integrativepediatricsonline.com/blog/2014/06/24/23andme-genetic-ttesting-dr-paul-recommends-and-explains-its-use/>.

Article Selection and Analysis



Duke Center for Health Informatics

Article Selection and Analysis

- Collected in Excel
- Read first 50 abstracts and titles
- 11 themes identified
- Created matrix
- Read 34 → 18
- No statistical analyses

Objective: investigate prevalence of info regarding techniques used for EHR data

Analysis

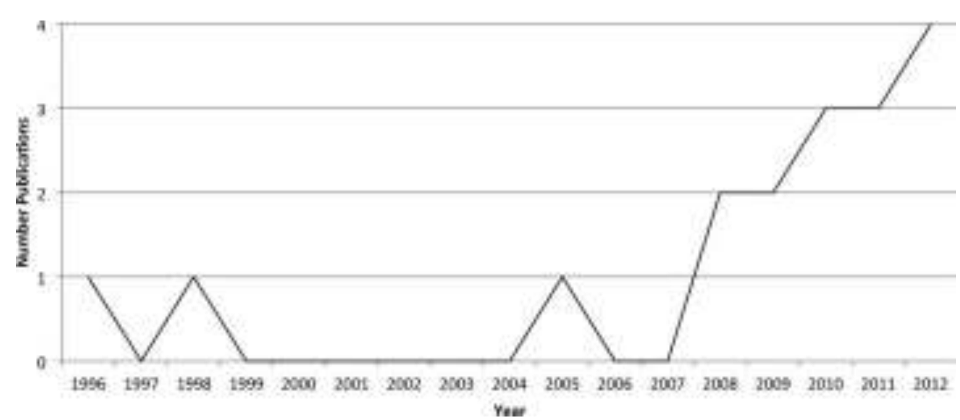
No.	Visualization Discussed	Use
4	LifeLines[11, 14-16]	Most advanced application; provides timeline of a patient's temporal events
1	LifeLines2 [19]	Enables use of multiple patient records; users see both numerical and categorical data; evolves into LifeFlow [31], used for millions of patient records to understand trends
4	VISITORS [17-20] (evolved from KNAVE/KNAVEII)[12,18]	Evolved from KNAVE/KNAVEII to accommodate diverse temporal data from multiple records; system feasible for exploring longitudinal data but interface needs simplification
1	Radial Starburst [34]	Shows complexity of data represented over 100-dimensional space; has potential for interactive system

Analysis

No.	Visualization Discussed	Usability
1	Pattern Matching & Temporal Interpolation[29]	Might have been excluded; used EHR data for 3 million patients; showed value in using time in correlation & aggregated data from many records vs a single record
1	DICON (Dynamic Icon)[30]	Interactively explore clusters of similar patients; clusters are represented as icons on a treemap; unique, but requires time for users to understand
1	Outflow [32,33]	Looks at disease progression paths; allows users to look at a visual display consisting of multiple events, their sequences, and outcomes
5	Other: ie. Treemaps,[28,30] radial displays, [34 36 38] icicle trees.[31]	

Analysis (n=18)

- 16 focus on use of visualizations for clinical decision support;
- Other 2 suggest use for quality assurance and improvement
- 15 studies address the use of temporal data
- Most describe interactive visualizations
- Reported training time ranged from 6 to 30 minutes



Common Themes & Challenges

- Themes:
  - Type of data accessible to the user
  - Meaningfulness of visualizing large amounts of data
  - Usability
  - Training time
- Challenges:
  - Data (quality, size, diversity)
  - Users (needs, skills)
  - Design (ability to visually explore & analyze results)
  - Technology (tools, infrastructure)

Challenges

- The amount of EHR data and its display is a challenge
  - Difficult to see and identify meaningful patterns in visualizations
  - Zoom, pan, and filter tools reduce clutter but will not suffice for 'big data'
- Size & complexity of EHR data is a challenge
  - Color, density, and filtering techniques distinguish variables
  - No reported techniques discuss applicability to entire datasets from EHR and potential for knowledge discovery
- Ability to use temporal data in visualizing aggregate data from EHRs is important to users
- Need design that presents a single interactive screen

Challenges

- Awareness of many variables that can lead to uncertain data in EHRs, potentially distorting temporal events
- Complications from missing values, inaccurate data, and mixed data types
- Users want to see categorical and numerical data both on a large scale and in detail
- Normalization scheme for aggregated numerical data
- Training time considerations, complexity of the data

Limitations of Study

- Review limited to primary publications describing innovative visualization techniques and application to EHRs
- Intentionally broad search terms
- Eliminated articles with more technical abstracts and a focus on geospatial representation
- No books

Conclusions

- Few techniques effectively and efficiently display large and complex data in EHRs
- Need techniques to handle Big Data & temporal data
- Look to techniques from other disciplines
- Research to date has identified important findings that can help guide future research

Funding

Supported by the US Army Medical Research and Material Command (USAMRMC) grant number W81XWH-13-1-0061

Duke Center for Health Informatics

Appendix J

Visualization of the Healthcare Visualization Literature

Vivian L. West, PhD, MBA, RN[1], David Borland, PhD[2], David A. West, PhD[3], W. Ed Hammond, PhD[1]
[1]Duke University, Durham, NC; [2]RENCI, The University of North Carolina at Chapel Hill, NC; [3]East Carolina University, Greenville, NC

Abstract
Text mining has been used in a variety of applications to discover information in unstructured textual data. Information visualization can help users see things in data that would otherwise not be evident. We combine the two techniques to better understand what the healthcare visualization literature contains.

Introduction and Background
Information visualization as a means to help users of very large data sets understand what is in the data is of increasing interest in healthcare researchers eager to use this big data to discover ways to improve healthcare outcomes. With any research project using visualization, researchers will seek knowledge from published research on its use. A literature search in MEDLINE or PubMed returns thousands of published articles, surprising since information visualization is a fairly recent approach in healthcare research.

Visualization, information visualization, and data visualization are not recognized terms in the lexicon of the National Library of Medicine's Medical Subject Headings (MeSH). Of the volume of articles returned in a literature search using any of the terms, the focus is quite diverse, leaving the researcher with the daunting task of manually reviewing the articles to select those of relevance. Text mining helps this process by using natural language processing (NLP) to provide the researcher with key words or terms in the unstructured data. With thousands of articles and the number of terms NLP identifies for each, the task is better but still overwhelming. Representing key words and terms visually can quickly help users identify the most frequently used terms, relationships, and correlations, and address the problem of information overload. The purpose of this research is evaluate the effectiveness of using text mining and visualization to understand what has been reported in the healthcare visualization literature.

Methods
The PDF files of 70 articles previously retrieved from a systematic review of the healthcare visualization literature[1] were imported into SAS Text Miner. NLP was used to parse each document and its constituent words and terms. One of four different frequency metrics identified a set of key terms (number determined by user, from 2,000 to 5,000) that efficiently discriminate between the documents. Singular value decomposition converted the weighted term matrix into a numerical vector for each document. The document vector was used for distance calculations and the formation of clusters and to form four higher-order themes for the document population.

After converting text to numerical data, visualization techniques are applied to the terms and concepts. Interactive visualization methods, such as force-directed network visualization, enable users to visualize multiple entities with the same values. This approach makes the information from large numbers of documents quickly understandable.

Results and Discussion
Text mining enables us to identify the clusters and concepts from the healthcare visualization literature. Visualization of these clusters and concepts further enables us to determine those prevalent in the literature and those missing, and provides a method for understanding numerous terms and their relationships with each other and where further research is needed, such as when terms are missing or lack prominence.

Conclusion
Text mining the unstructured data in many publications about information visualization in healthcare and visualizing the results is a powerful way to understand what this literature contains. We are now working with a much larger data set. The same techniques can be applied to any large body of unstructured data, leading to discoveries that have the potential to drive research in ways not previously thought of.

References

1. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: A systematic review. JAMIA. Forthcoming 2014.

Appendix K and L

The Use of Data Visualization in Transforming Care Delivery

Vivian West, PhD, MBA, RN; Duke Center for Health Informatics

David Borland, PhD; Renaissance Computing Institute, The University of North Carolina at Chapel Hill
W. Ed Hammond, PhD; Duke Center for Health Informatics
Eugenia McPeek Hinz, MD; Duke Health Technology Solutions
Igor Akushevich, PhD; Duke University

NCHICA 21st Annual Conference; Pinehurst, NC
September 14, 2015

What is Visualization?

The question is not what you look at, but what you see." (Henry David Thoreau, 1851)

Many an object is not seen, though it falls within the range of our visual ray, because it does not come within the range of our intellectual ray, i.e. we are not looking for it. So, in the largest sense, we find only the world we look for. (Henry David Thoreau, 1857)

We cannot see anything unless we are possessed with the idea of it, and then we can hardly see anything else. (Henry David Thoreau, 1858)

What is *Information* Visualization?

"Using Vision to Think"

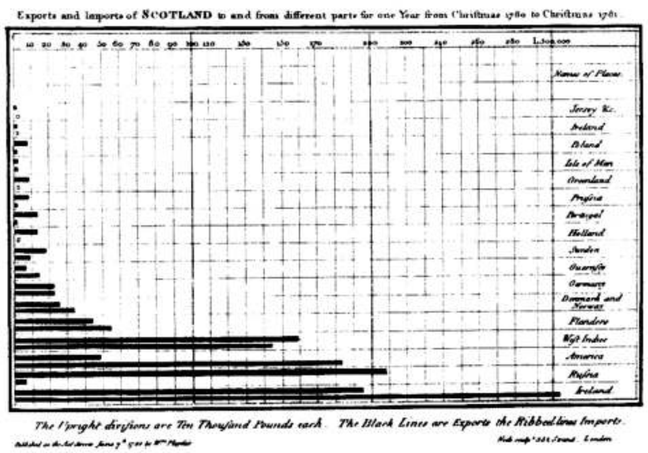
(Readings in Information Visualization: Using Vision to Think; Card, Mackinlay, Shneiderman; 1999)

"The use of computer-supported,
interactive,
visual representations of abstract data
to amplify cognition."

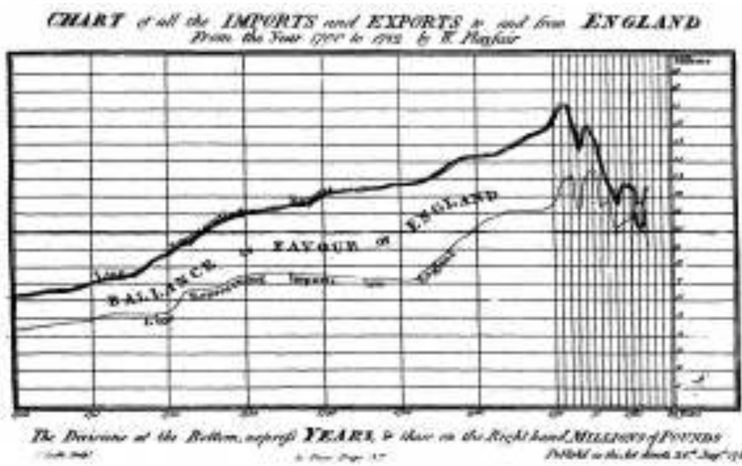
Duke Center for Health Informatics

Historical Perspective of Visualization

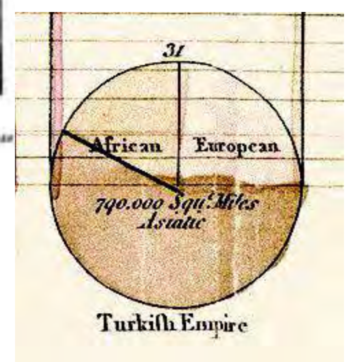
William Playfair: 1700s



One year trade data for Scotland
17 trading partners exports and imports
From Commercial and Political Atlas (1786)



Time-series line graph
England 18th century imports/exports
From Commercial and Political Atlas (1786)

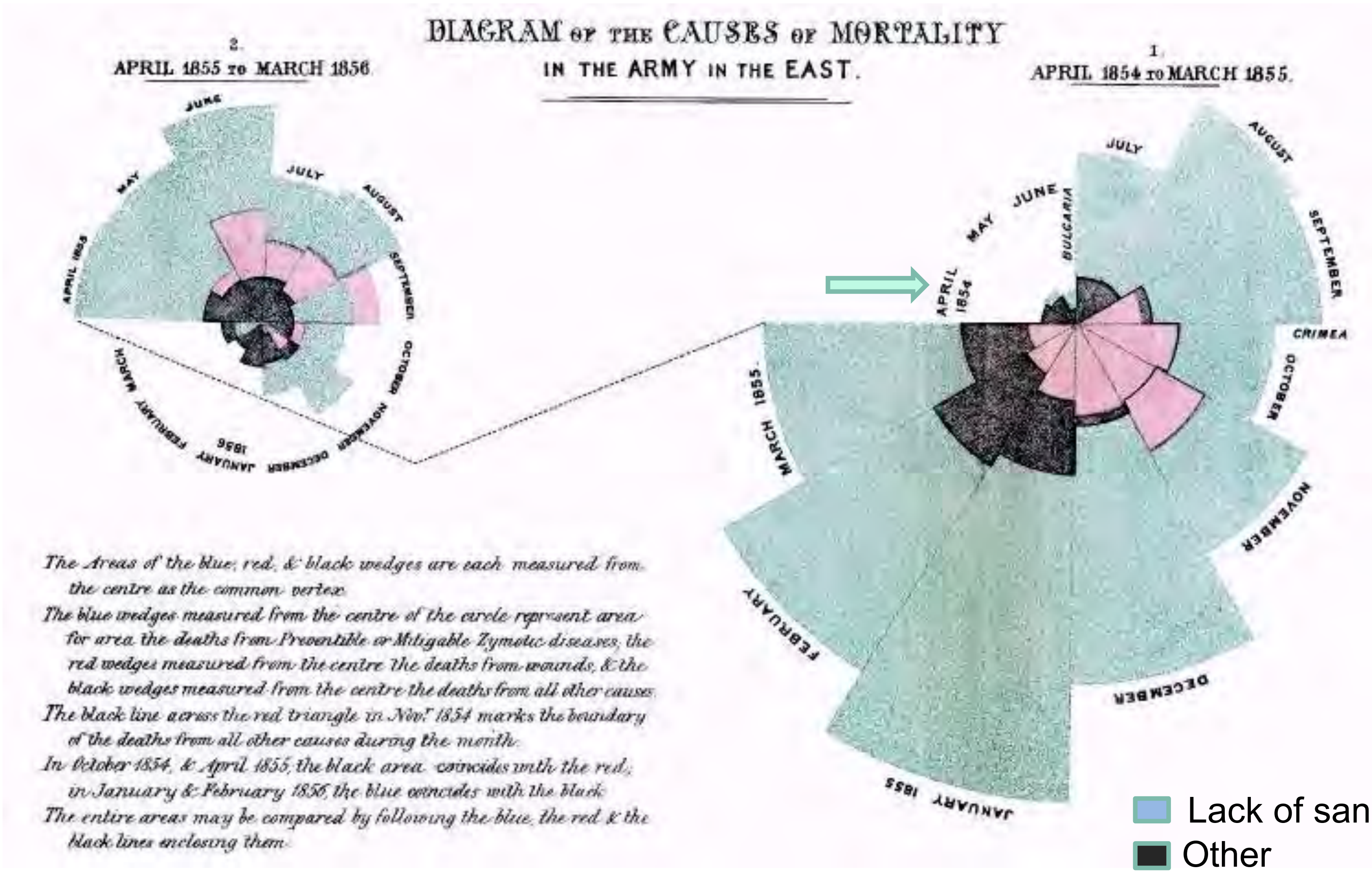


Turkish Empire: proportions located in Africa, Europe, Asia before 1789
From Statistical Breviary(1801)

Historical Perspective

Florence Nightingale

Coxcomb chart 1854-1856 Crimean War

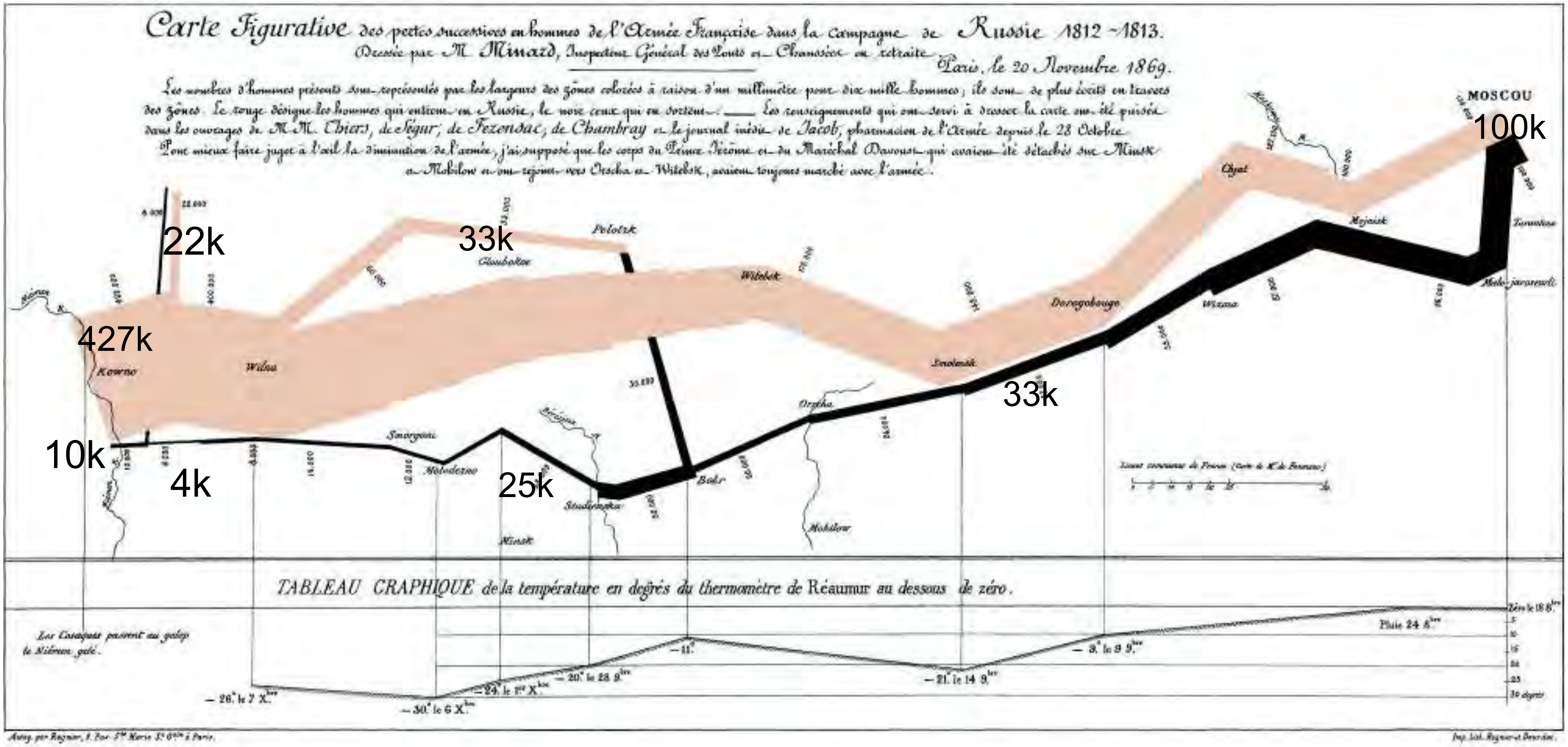


- Lack of sanitation
- Other
- Wounds from war

Historical Perspective

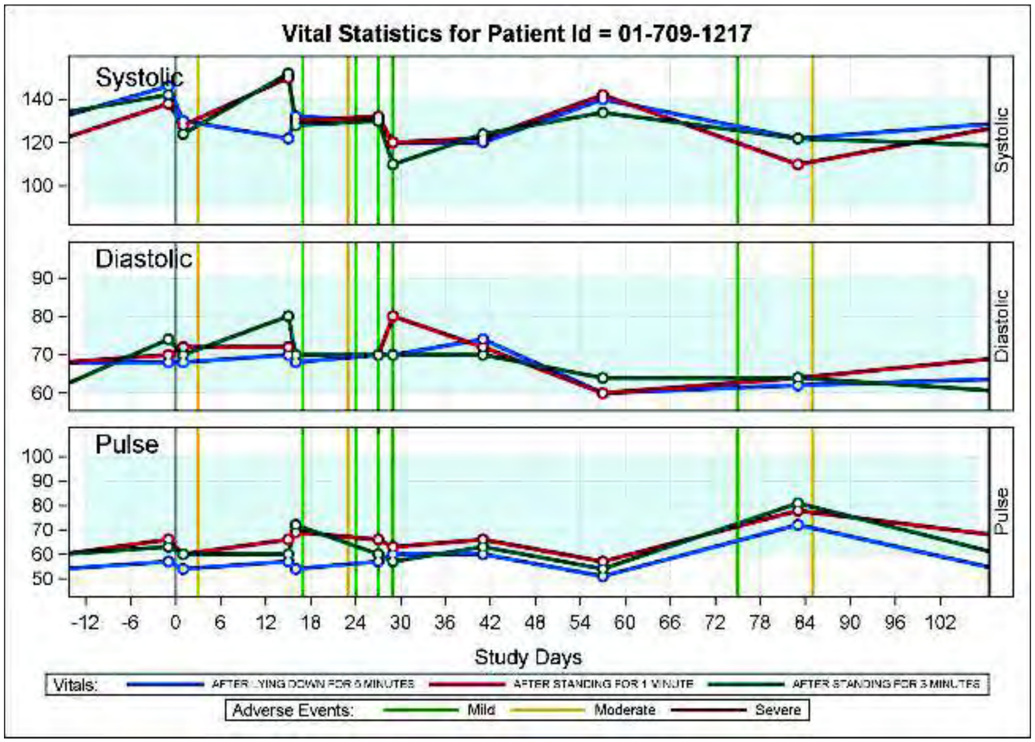
Charles Minard 1869

Loss of men in French army in Russian campaign 1812-1813



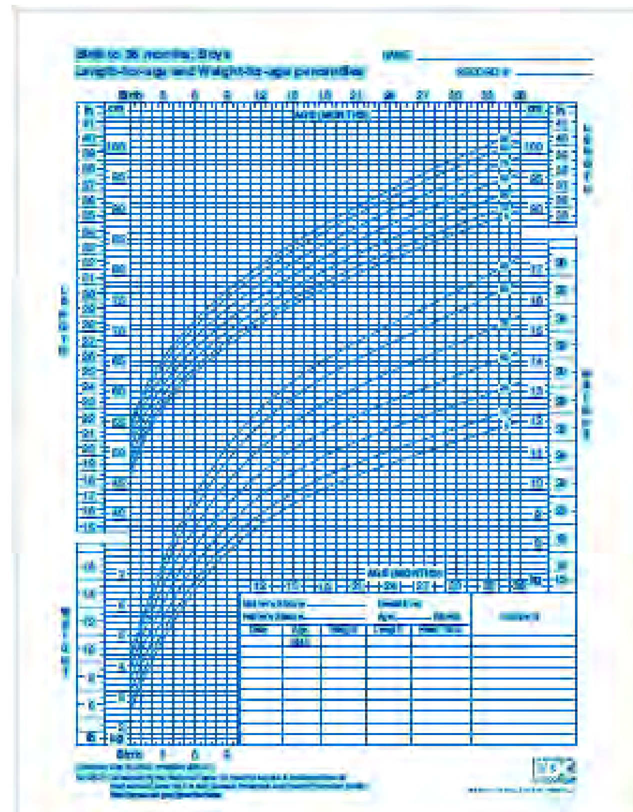
Duke Center for Health Informatics

Visualization in Health Care



Matagne S. Patient Profile Graphs Using SAS, Paper 160-2013.

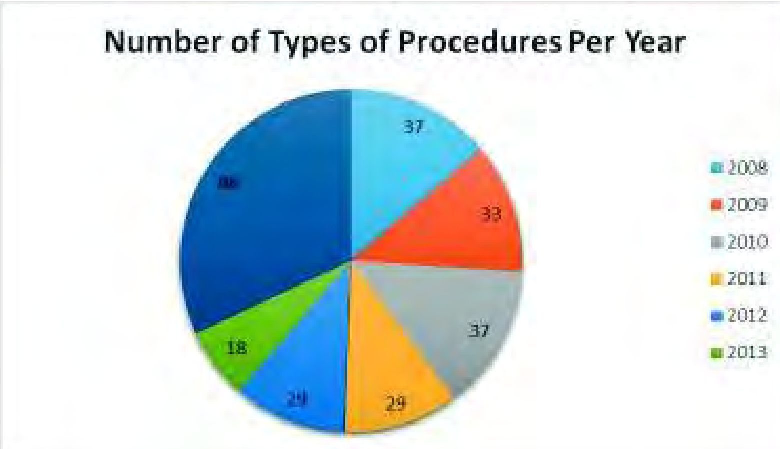
CAP, Electronic Laboratory Data Exchange

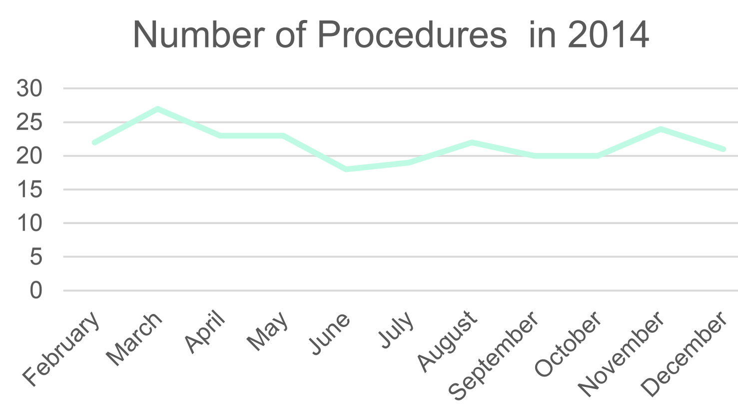


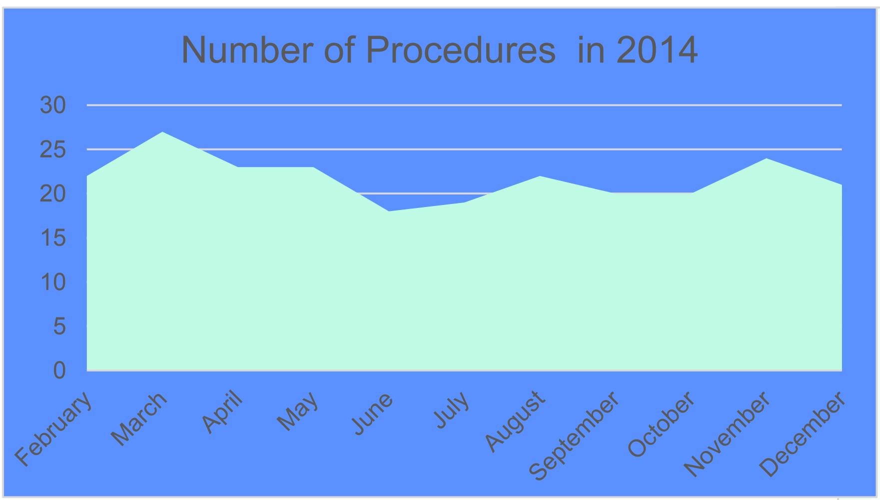
CDC, May 30, 2000

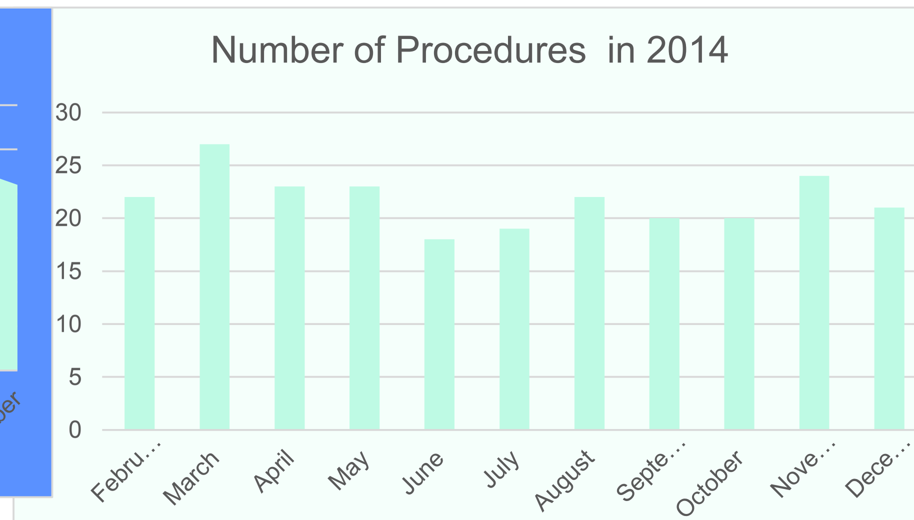
Duke Center for Health Informatics

Examples

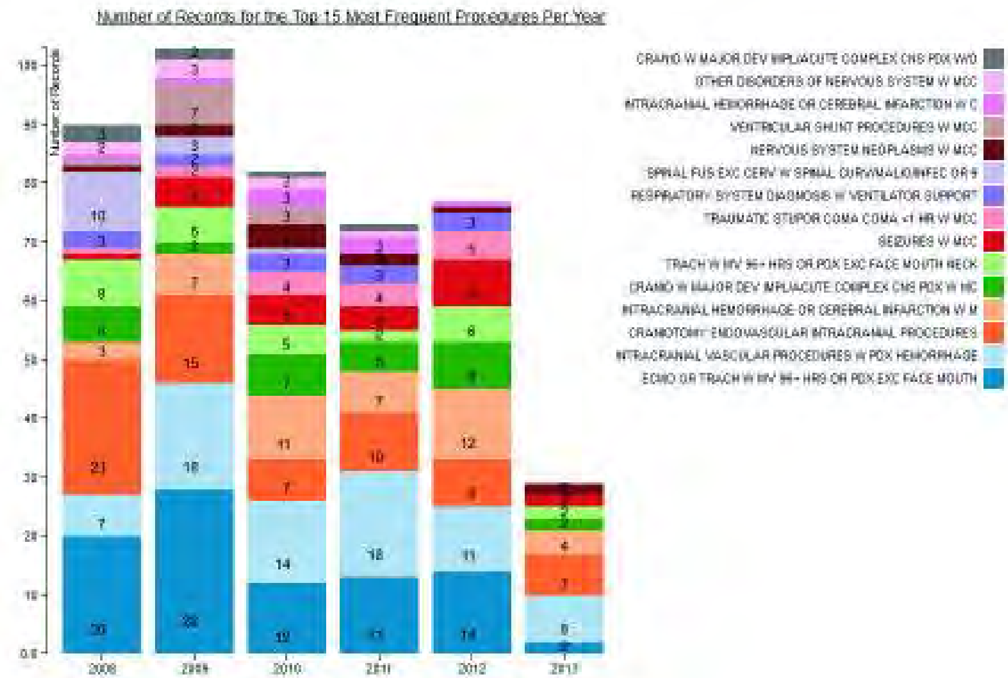


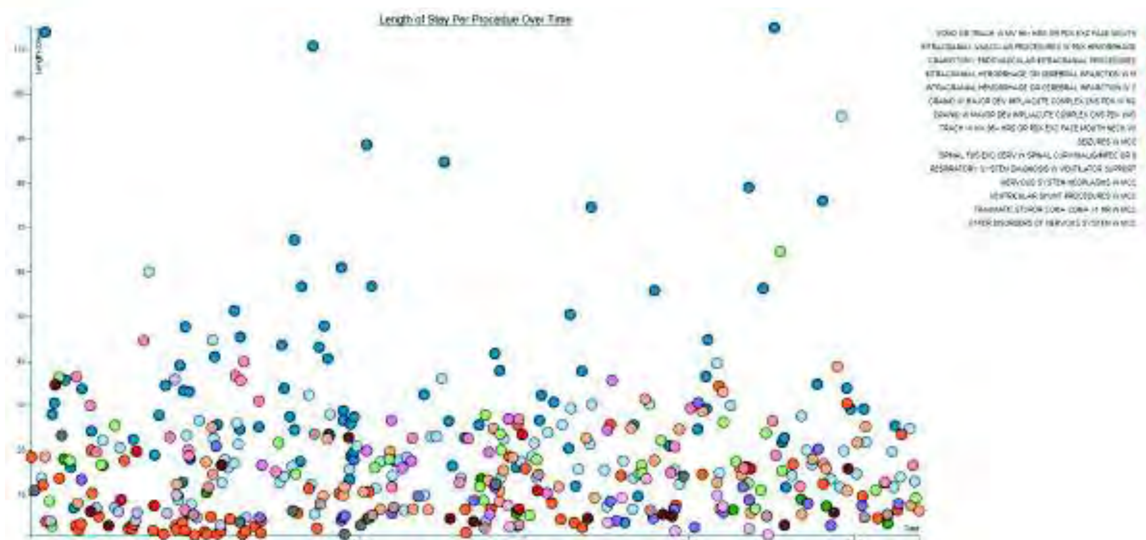






Duke Center for Health Informatics





All examples courtesy of
Leigh Ann Herhold.

Duke Center for Health Informatics

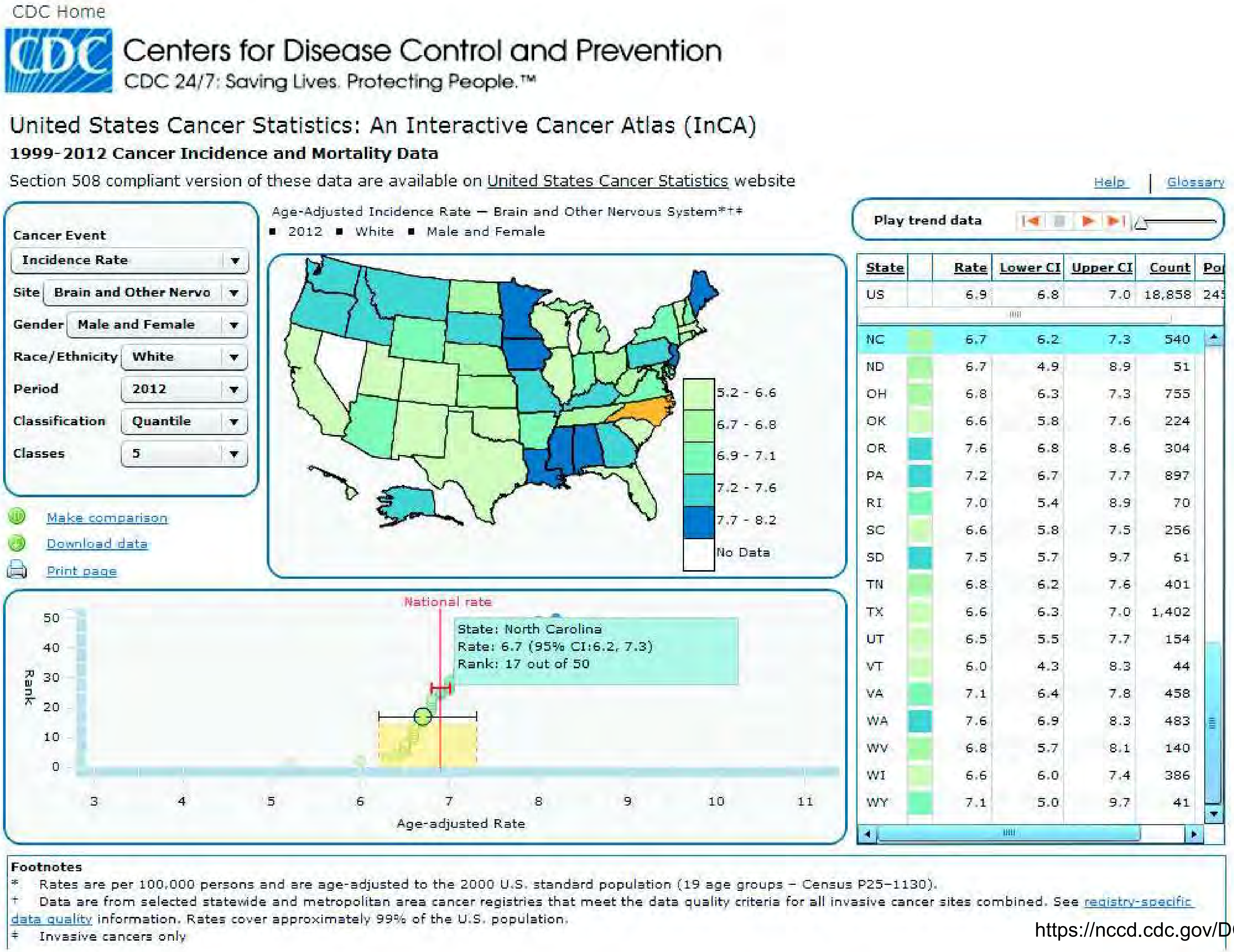
Interactive Visualization

- Human input
- Improve presentation and understand data
  - Choose options from a list
  - Color changes
  - Zoom in on point of interest
  - Expand (drill down) on information

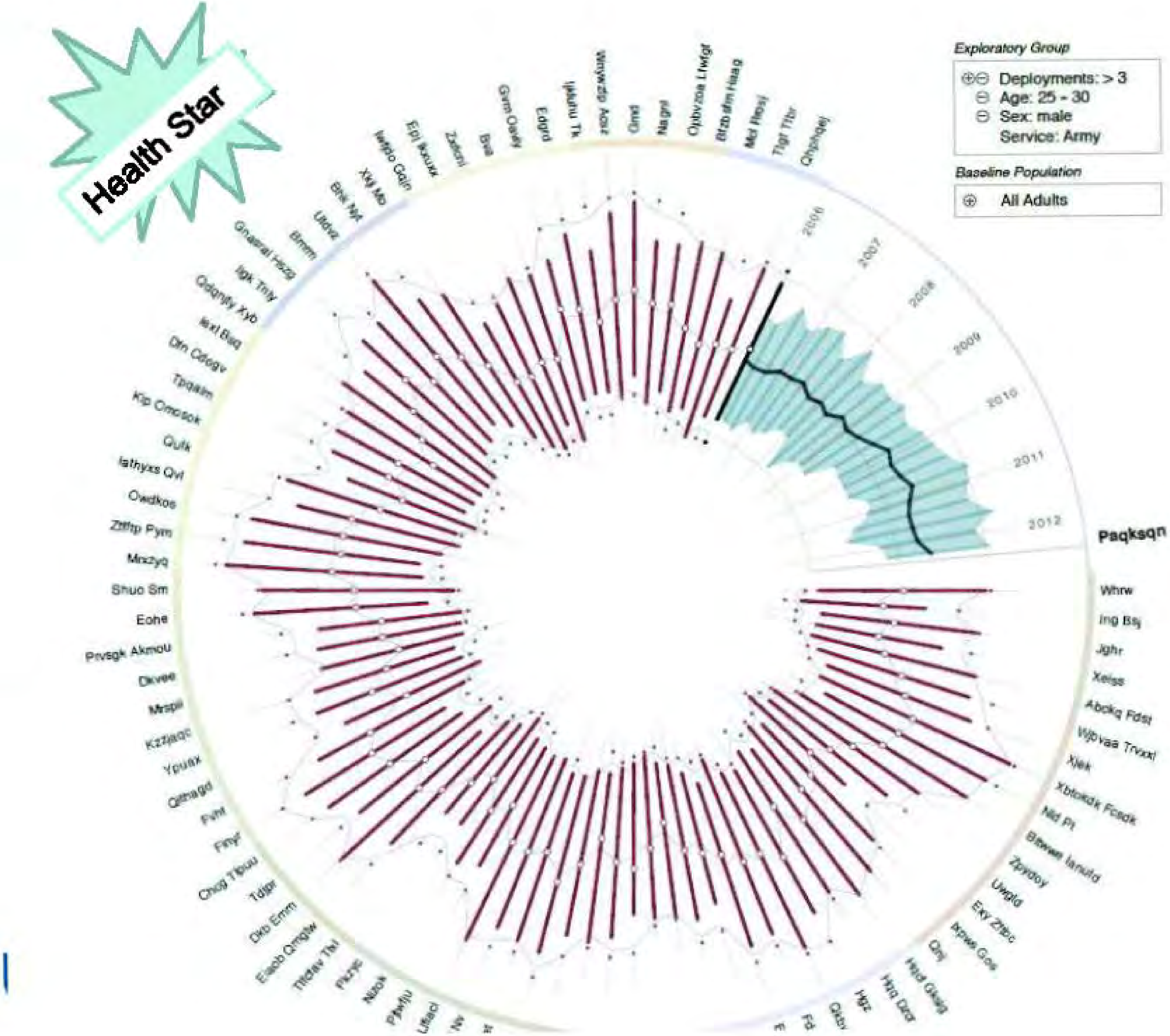
Use of information visualization

- Finance
- Accounting
- Petroleum industry
- Engineering
- Genetics
- CDC maps, timelines

CDC Interactive Atlas



https://nccd.cdc.gov/DCPC_INCA/

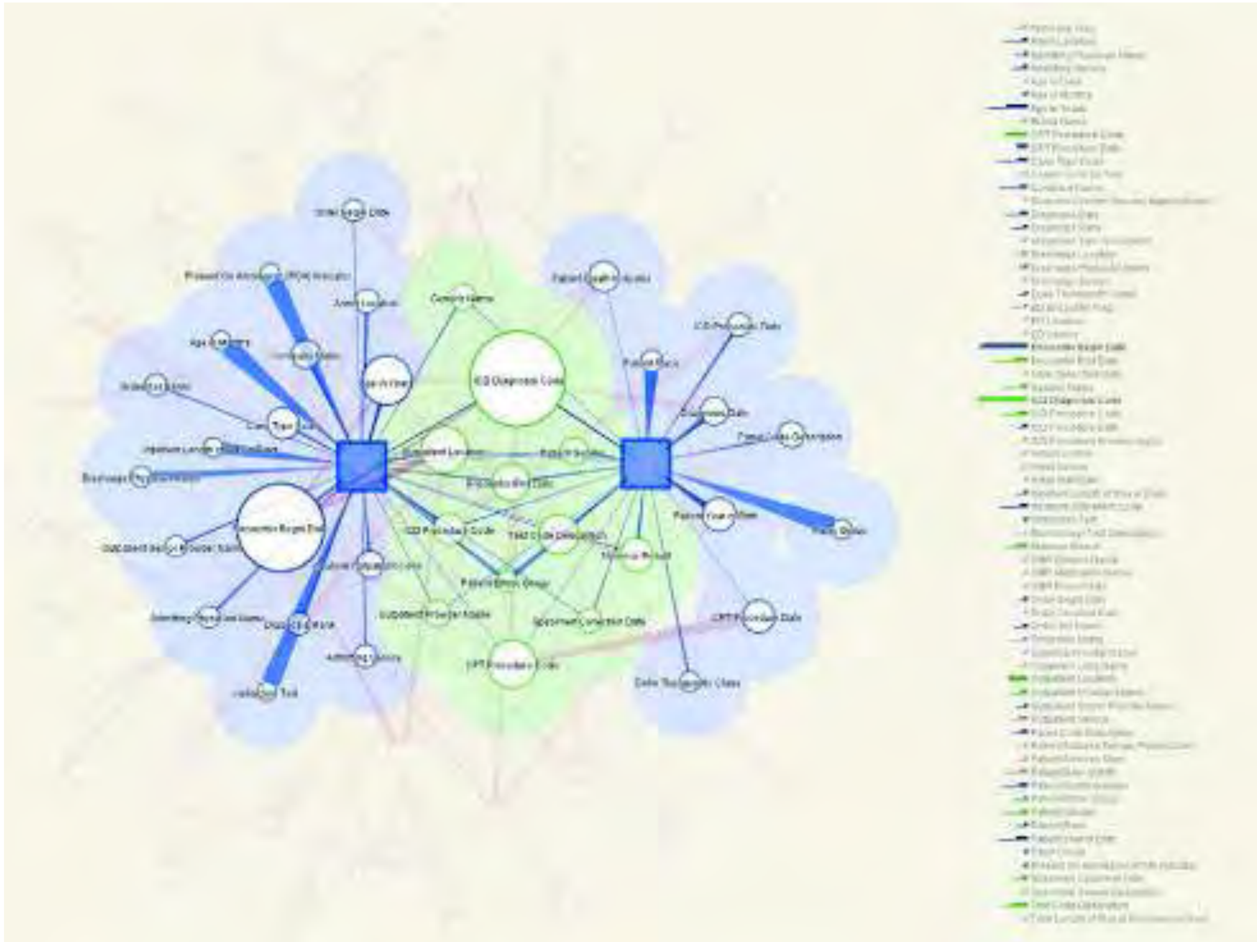


Health Star
Exploratory Group
Deployments: > 3
Age: 25 - 30
Sex: male
Service: Army
Baseline Population
All Adults
Paqksqn
2006
2007
2008
2009
2010
2011
2012
Whrw
Jghr
Xelss
Abckq Fdst
Shuo Sm
Eohe
Prvsgk Akmou
Owdkos
Mrszyq

DoD Exploratory Study

- What information and knowledge are in EHRs?
- Explore interactive visualization of large sets of health data to better understand what is in the data.
  - Duke EHR data
  - BT data
  - Simulated data from DoD
- February 25, 2013 – August 24, 2015

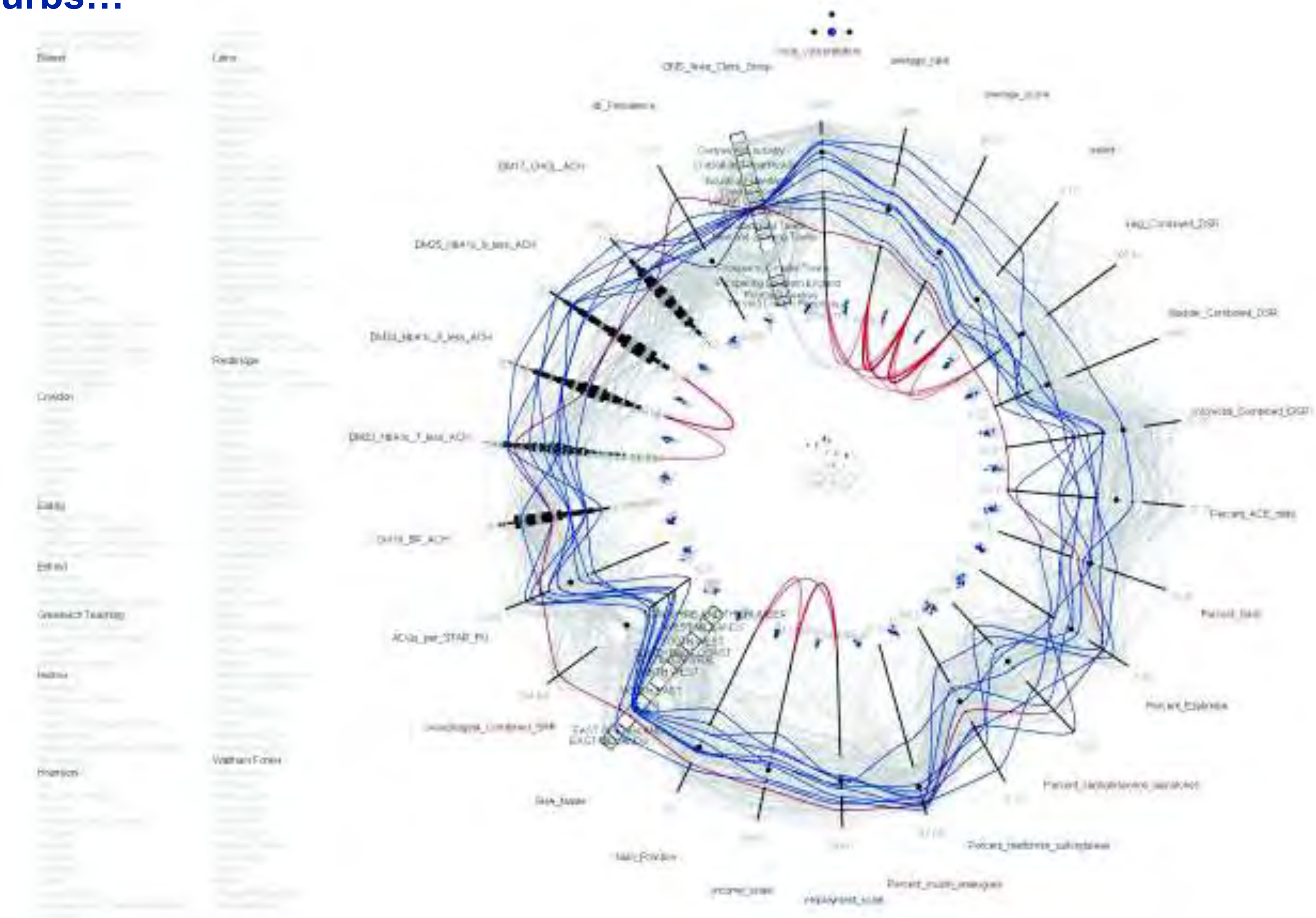
DEDUCE Data



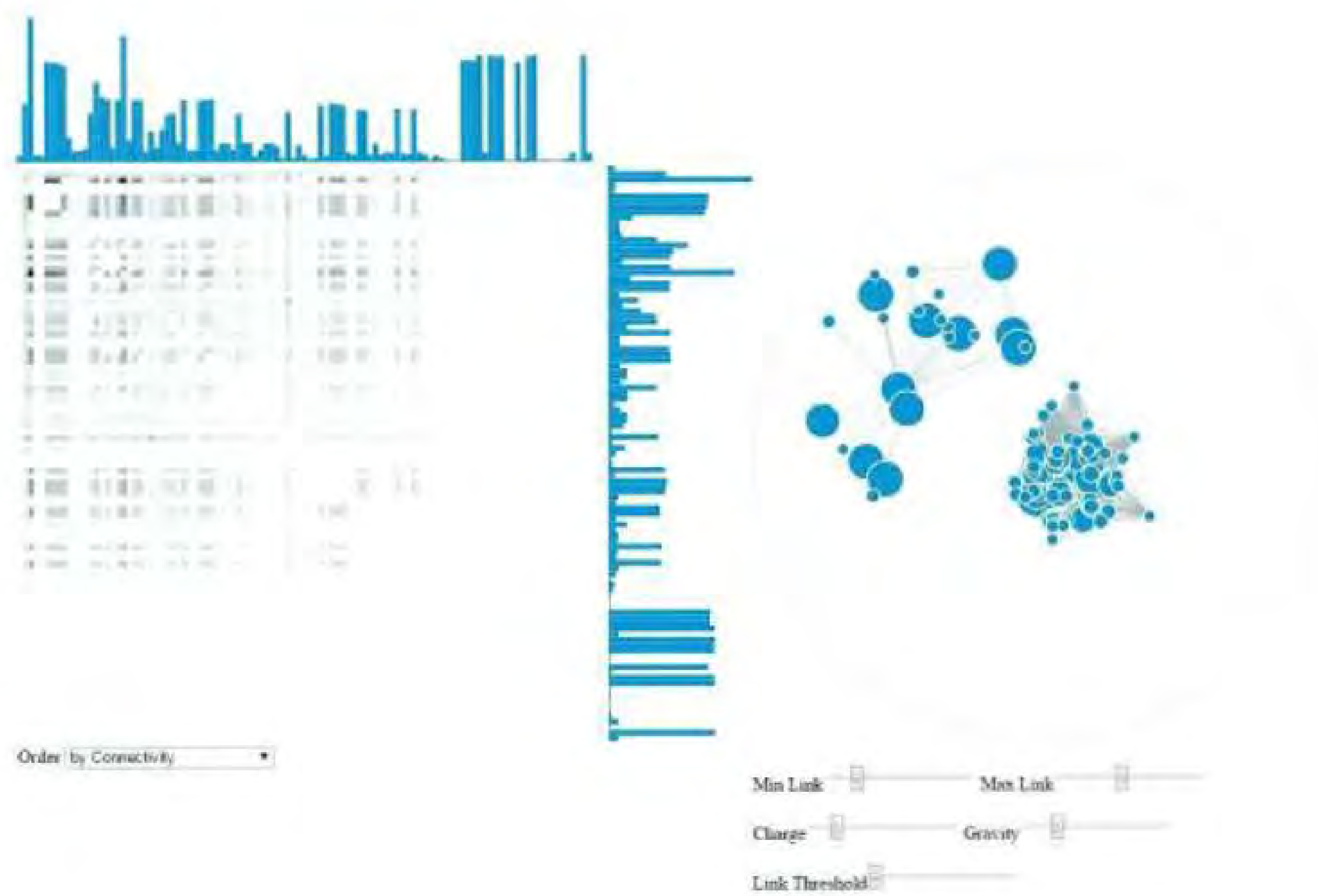
Duke Center for Health Informatics

BTS Visualization

Harrow vs. other London Suburbs…

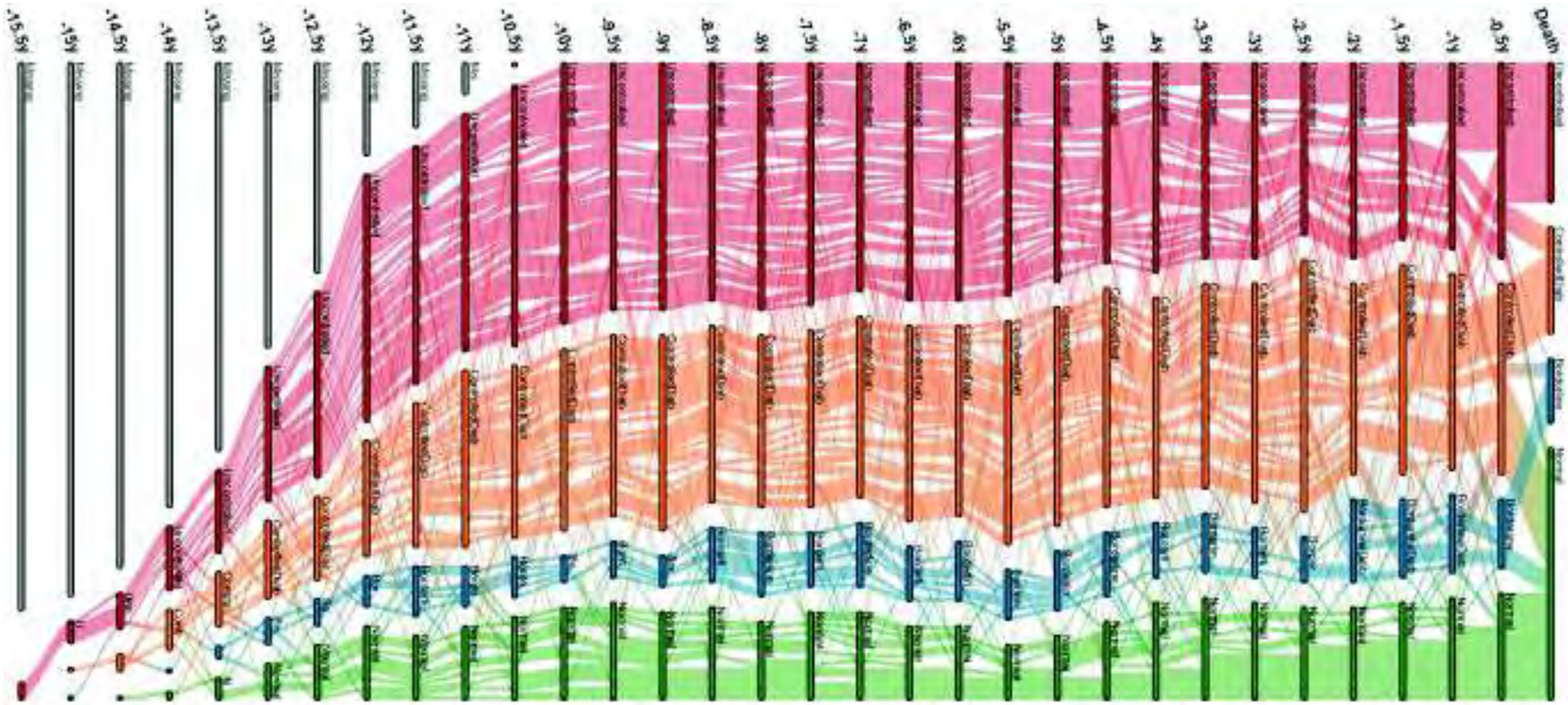


AHLTA Data vs Simulated Data

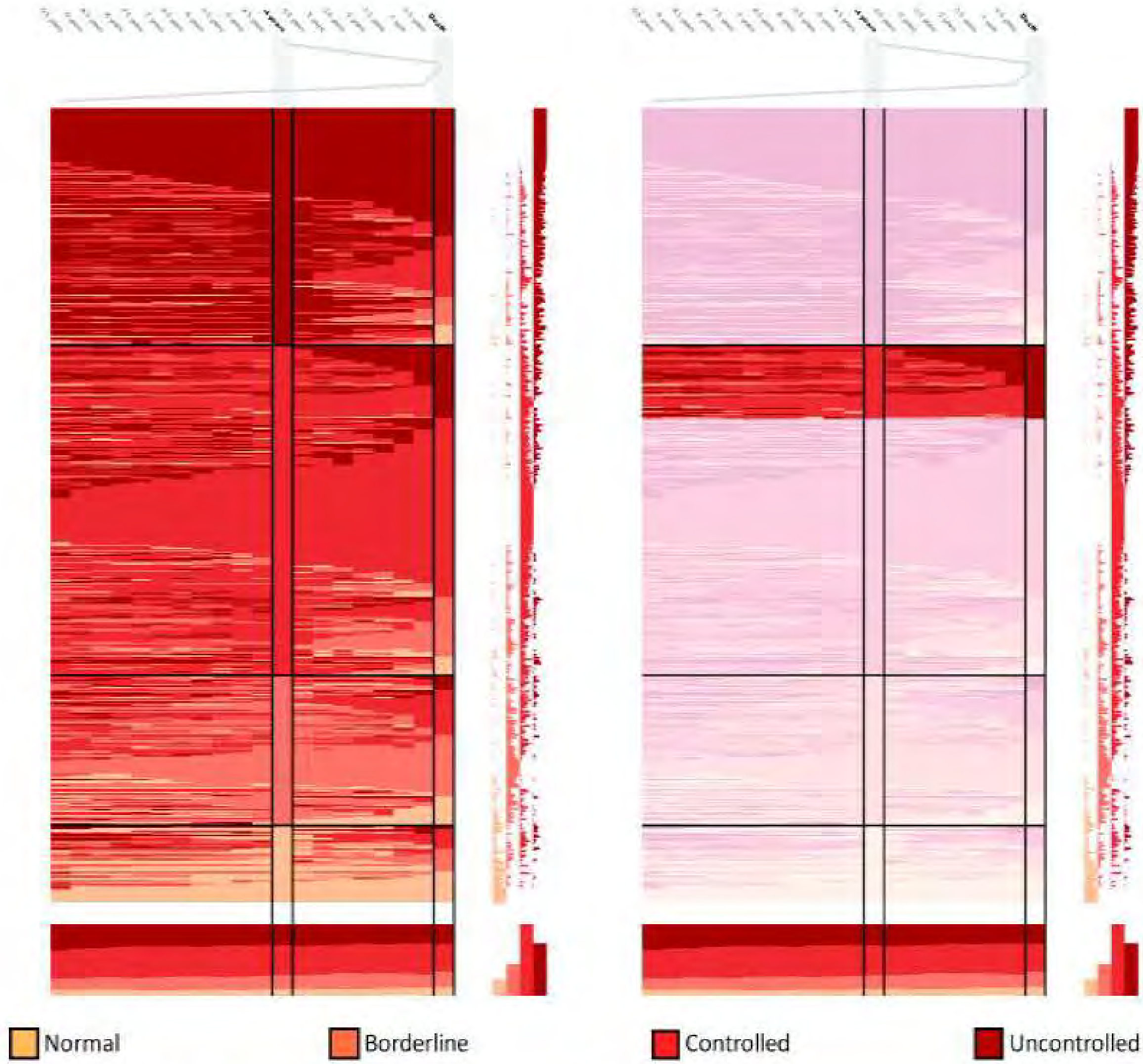


Duke Center for Health Informatics

T2D Visualizations

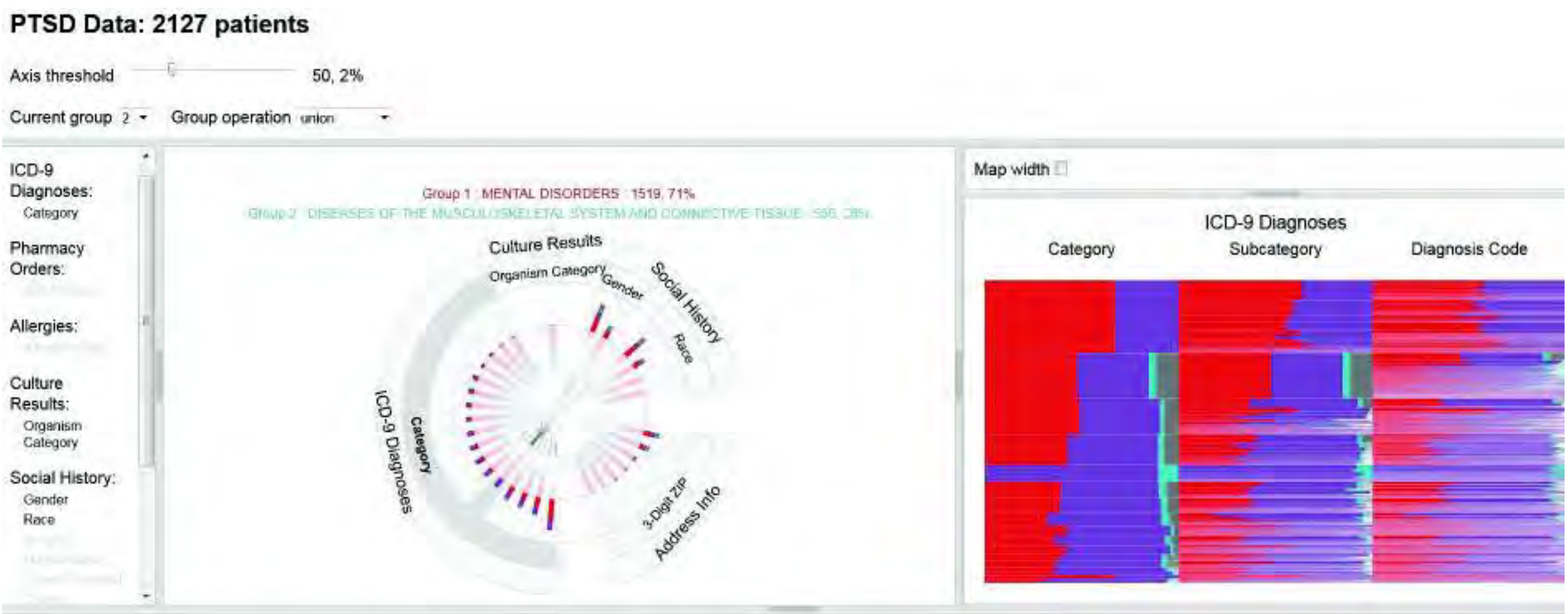


PathMap (n=3,638)



Duke Center for Health Informatics

PTSD Visualization



Best Visualizations for Specific Data Elements

- Confounding variables
  - Age
  - Work experience
  - Individuals vs populations
  - Experience seeing information visually
  - Orientation to facts versus trends

Used to looking at pictures and for patterns in what is seen.

Challenges of Information Visualization

- Usability
  - Legends
  - Labels
  - Color
  - Size of graphical representations
  - Placement of axes
  - Ease of understanding at a glance
  - Amount of data represented
  - Size of display

Challenges of Information Visualization

- Manage massive amounts of data
- Display temporal data
- Complexity of data
- Display categorical and numerical data
- Manage “clutter”
- Missing data
- Inaccurate data entry
- Training time

Normalization scheme

Research Avenues

"We cannot see anything unless we are possessed with the idea of it, and then we can hardly see anything else."

- Numerous disease-specific applications
- Population health

Acknowledgements

Work supported by the US Army Medical Research and Material Command (USARMC) under Grant No. W81XWH-13-0061. The views, opinions and/or findings contained in this presentation/report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

Thank you!

Questions?

Appendix M

Path Maps: Visualization of Trajectories in Large-Scale Temporal Data

Category: Research

ABSTRACT

Understanding how a given quantity changes over time for multiple entities is a common task when analyzing time-varying data sets. Various temporal visualization techniques exist, however many of these techniques are ineffective for large data sets. We introduce the path map, a temporal visualization technique for regularly-sampled ordinal data, designed to effectively handle data sets with many entities. The path map is a rectangular space with columns representing temporal samples and rows representing individual data entities. Rectangular cells with a single color-mapped value are generated from combinations of adjacent rows based on their vertical ordering. An interactive sorting interface effectively organizes and reveals patterns in the data by reordering the position of each row based on its values at user-selected columns. Additional contributions include missing data display and aggregation methods to handle larger data sets. We demonstrate path maps with lab data from over 500 and over 3500 diabetic patients, taken over a period of up to ten years before death.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces; I.3.8 [Computing Methodologies]: Computer Graphics—Applications; J.3 [Computer Applications]: Life and Medical Sciences—Health

1 INTRODUCTION

Various visualization techniques for time-varying data, such as line graphs, stacked graphs, and horizon graphs, effectively show how a given quantity changes over time for various entities. However, when dealing with large data sets consisting of hundreds or thousands of entities, the effectiveness of such techniques is reduced due to problems such as over-plotting.

We are studying the temporal trajectories of measures, such as lab values, related to various diseases for large cohorts of patients in electronic health record (EHR) databases. Many temporal visualization techniques developed for health-related data, although effective, specifically target sequences of discrete events [3, 4, 5], and are not directly applicable to our problem. More closely related is [2], which also shows trajectories in time-varying clinical data. While effective for small numbers of temporal samples, the visual complexity increases with the number of samples, making the resulting visualization difficult to interpret when there is a lot of variation in the data. Reducing the number of samples hides any variability between the remaining samples.

We have therefore developed the path map, a visualization technique designed to reveal patterns in large temporal data sets. Path maps are based on heat maps, similar to [1], however this technique was applied to relatively small data sets. The path map is a rectangular space with columns representing temporal samples and rows representing individual data entities. Rectangular cells with a single color-mapped value are generated from combinations of adjacent rows based on their vertical ordering. An interactive sorting interface enables the user to effectively organize and reveal patterns in the temporal data by reordering the vertical placement of each row based on its values at user-selected temporal samples, while still showing the variability between selected samples.

Additional contributions include the missing data display and data aggregation methods to reveal patterns in larger data sets. We demonstrate path maps with lab data from over 500 and over 3500 diabetic patients, taken over a period of ten years before death.

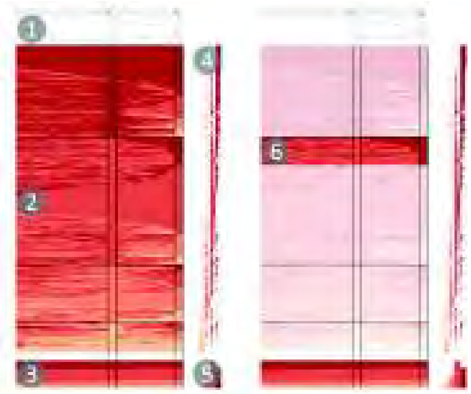


Figure 1: Path map of 546 patients with contiguous map method and custom sorting: 1) sorting interface, 2) path map, 3) column overview, 4) row overview, 5) value overview, 6) highlighting all patients who moved from controlled (orange) to uncontrolled (red) between -4 years and Death.

2 METHODS

2.1 Data Description

The path map assumes ordinal data with regular temporal samples. The data shown here are hemoglobin A1c (HbA1c) levels from diabetic patients, aligned by date of death on the right. Visualization tasks of interest include showing general trends, e.g. whether HbA1c levels tend to increase or decrease over time, and identifying subpopulations of interest, e.g. groups of patients that maintain elevated HbA1c levels or whose levels normalize before death. We compute average HbA1c every six months, starting ten years from death. When there are no readings in a sample interval, the previous value is carried forward, and the sample marked as missing. Samples prior to the first reading for a patient are colored gray. HbA1c values are categorized according to clinical guidelines as normal, borderline, controlled, or uncontrolled, and color-mapped from yellow to red.

2.2 Basic Path Map

The basic path map visualization represents each data entity as a row and each temporal sample as a column (Figure 1). Benefits of the path map representation include ensuring no line crossings, making each individual trajectory easier to follow than with standard line graphs in the presence of many data entities, and an information-dense display of many temporal data entities. A *contiguous* map method is used, which generates cells from rows with vertically contiguous values per column. An overlay highlights selected columns and outlines regions based on the primary column. The path map visualization also provides row, column, and value overview visualizations to show general trends in the data. Section 2.4 describes different map methods to handle large data sets. For all map methods, highlighting any cell shows the distribution of that cell's rows in all other cells.

2.2.1 Sorting Interface

The vertical position of each row is crucial for finding patterns of interest. We provide five basic sorting methods: *weighted aver-*

Figure 2: Sorting interface with custom sorting by value at -6 years, then -2 years, then Death, then backwards from Death.

age, *forward*, *backward*, *first*, and *last*. For each, the user selects a column of interest, and each row is sorted by its value at this column, resulting in a stacked bar. Within each group of rows with the same value, the *weighted average* method sorts by weighted average around the selected column, *forward* and *backward* sort by the value at each subsequent column in the indicated direction (reversing direction at the last column in that direction), and *first* and *last* sort by the value at the indicated column, then forward or backward from there. The *last* method is especially useful for our purposes, as it shows the breakdown of HbA1c values at death given the HbA1c values at the selected sample before death (Figure 1).

A *custom* sorting method enables the selection of any number of columns to sort by, in any order, and a sort direction (forward or backward) for sorting by unselected columns. The interactive sorting interface provides a visualization of the sort order for the basic and custom sorting methods (Figure 2).

2.3 Data Status Display

Our data preprocessing carries forward the previous value when there is missing data for a sample interval. As it may be useful to know whether a given data point is missing or not, we provide the capability to display data status, such as missing, via a striped pattern. Two sorting options are available: sorting by status *first*, or *second*. Sorting *first* more effectively shows the total amount of missing data, whereas sorting *second* more effectively shows the breakdown of missing data per data value (Figure 3).

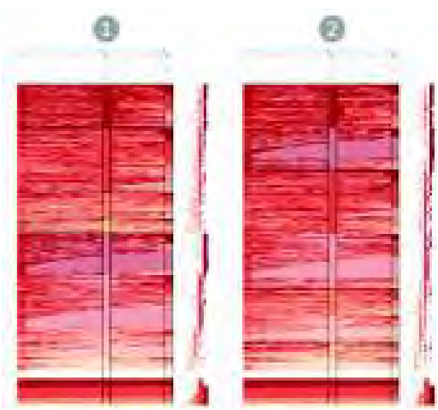


Figure 3: Missing data status display via striped pattern. 1) Sorting by status first emphasizes the total amount of missing data. 2) Sorting by status second emphasizes the amount per data value.

2.4 Data Aggregation Map Methods

Whereas the *contiguous* map method works well for moderately-sized data sets, over-plotting becomes a problem when the number of rows exceeds the number of vertical pixels in the path map. We have therefore developed map methods that aggregate data to more effectively handle larger data sets (Figure 4). The same sorting interface is used, however between selected columns aggregation methods are used to generate cells. Rows can be grouped for aggregating based on common values at the column earlier in the sort order (larger groups), or later in the sort order (smaller groups).

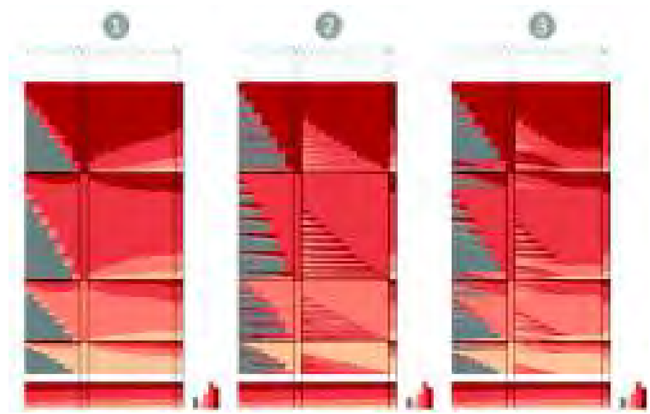


Figure 4: Data aggregation map methods for 3638 patients: 1) column summary, 2) row hierarchy, 3) row compress.

2.4.1 Column Summary

The *column summary* map method rearranges row samples to display stacked bar charts for each row group at each intermediate column. This method is most effective for showing general trends between values at the selected columns.

2.4.2 Row Hierarchy and Row Compression

The *row hierarchy* map method generates a hierarchical layout between each pair of selected columns based on value counts per row. The per-group hierarchy order is determined by the total value counts for the group. Cell width represents the count (number of samples with that value), and cell height represents the number of rows with that count for that value. The *row compression* map method is similar, however rows are initially grouped by the value with the highest count per row, then by remaining values. These layouts are useful for selecting groups of rows with certain characteristics, such as mostly uncontrolled values between two selected columns.

3 Conclusion

Our path map prototype has proven useful for the visualization of temporal trajectories of HbA1c values in thousands of diabetic patients. We are interested in exploring the use of path maps with other medical data, and incorporating additional patient data through coordinated views. We are also interested in using path maps with non-medical time-series data.

References

[1] L. Chittaro, C. Combi, and G. Trapasso. Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages and Computing*, 14(6):591–620, Dec. 2003.

[2] E. M. Hinz, D. Borland, H. Shah, V. L. West, and W. E. Hammond. Temporal visualization of diabetes mellitus via hemoglobin A1c levels. In *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare (VAHC 2014)*, 2014.

[3] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, 2013.

[4] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. Temporal summaries: supporting temporal categorical searching, aggregation and comparison. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1049–1056, 2009.

[5] K. Wongsuphasawat and D. Gotz. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2659–2668, Dec. 2012.

Appendix N

DECISION SCIENCES INSTITUTE

An Evaluation of Machine Learning Methods and Visualization of Results to Characterize Large Healthcare Document Collections

(Full Paper Submission)

Vivian L. West
Duke University, Center for Health Care Informatics
vivian.west@duke.edu

David Borland
Renaissance Computing Institute, The University of North Carolina Chapel Hill
borland@renci.org

David West
East Carolina University
westd@ecu.edu

W. Ed Hammond
Duke University, Center for Health Care Informatics
william.hammond@dm.duke.edu

ABSTRACT

This research is an exploratory analysis of the abilities of machine learning algorithms (namely text mining) and interactive visualization to analyze large collections of health care research documents. Preliminary results from the analysis of 391 documents describing research in health care information visualization are presented.

KEYWORDS: Healthcare, text mining, machine learning, data visualization

INTRODUCTION

In 2004 a presidential executive order, 'Electronic Health Records (EHRs) for All Americans', laid out tenets to improve the quality and efficiency of healthcare. One of its goals was to have accessible EHRs for most Americans within 10 years (Bush, 2004a; Bush, 2004b). In September 2009, years of research and policy work culminated in the Health Information Technology for Economic and Clinical Health Act (HITECH Act), allocating $19.2 billion in incentives to increase the use of EHRs by hospitals and health delivery practices. The Office of the National Coordinator (ONC) reports that at the end of 2014, 95% of acute care hospitals have EHRs that meet federal requirements for certification of EHR systems, and 75% of federal non-acute care hospitals have installed EHRs with core functionality (Charles et al, 2015).

With the burgeoning amount of electronic data, the field of biomedical informatics is also growing. In 1996 Ben Shneiderman aptly stated what remains true today: "Exploring information collections becomes increasingly difficult as the volume grows" (Shneiderman, 1996). There are calls for research using informatics tools and approaches to develop methods to extract and use this large amount of data, herein referred to as big data. With the significant potential for knowledge discovery, researchers are challenged to find innovative and effective ways to use the information from the rapidly growing big data now available from EHRs.

Graphical visualization is an effective tool that has been used since the later part of the 18th century to communicate information about data. A plethora of scales, shapes, and colors have been used with both small and large datasets rendered as visual diagrams such as bar charts, line graphs, scatterplots, and pie charts to reveal patterns leading to knowledge discovery. Industries such as finance, accounting, and the petroleum industry routinely use information visualization, defined as "interactive, visual representations of abstract data to amplify cognition" (Card et al, 1999) using innovative approaches that account for both the volume and complexity of the data. In the healthcare field, however, applications of advanced visualization techniques to large and complex datasets are limited.

We are currently working on a Department of Defense funded research project entitled Novel Visualization of Large Health Related Data Sets, exploring visualization techniques using big data from EHRs to discover what the data contain. With any research project using visualization, researchers will seek knowledge from published research on its use. Accordingly, we completed a systematic review of the visualization literature in May-June 2013 using primarily PubMed, the most frequently used reference database in health and indexed using the National Library of Medicine (NLM) controlled vocabulary (Medical Subject Headings, or MeSH) and Web of Knowledge. Of the volume of articles returned in the literature search, the focus was quite diverse, leaving us with the daunting and time-consuming task of manually reviewing the articles to select those of relevance. As interest in and publication about visualization of health related information increases, a tool to easily identify the topics covered by the literature would make this task more manageable.

The purpose of this research is to evaluate the effectiveness of using text mining and visualization techniques to explore and understand what has been reported in the health care visualization literature. With many articles and the large number of terms machine learning identifies for each, the ability to discriminate and identify significant articles related to our interests is better but still overwhelming. We hypothesize that representing key words and terms visually can quickly help users identify the most frequently used terms, relationships, and correlations, and address the problem of information overload.

LITERATURE REVIEW

Text mining in the field of biomedical informatics has become of great interest to researchers (Chaussabel, 2004; Hur et al, 2009; Shatkay & Feldman, 2003; Labaer, 2003; De Bruijn & Marin, 2002).Text mining biomedical literature is the topic for the March 2015 issue of Methods, a journal that focuses on experimental biological and medical sciences (Navarro & Iratxeta (eds.), 2015). The issue describes a sample of the text mining methods used in the field today. The concept of generating hypotheses from potential links in various publications using text mining, or literature-based discovery, has been used by a number of biomedical researchers (Srinivasan, 2004) to examine such topics as drugs (Androniz et al, 2011; Agarwak & Searls, 2008; Shetty & Dalal, 2011; Bellis et al, 2011), viruses (Hu et al, 2005; De Chassey et al, 2008; Their et al, 2012), and genetics (Hu et al, 2005; Papanikolaou et al, 2014; Poos, 2014; Xiang et al, 2013; Jung et al, 2014).

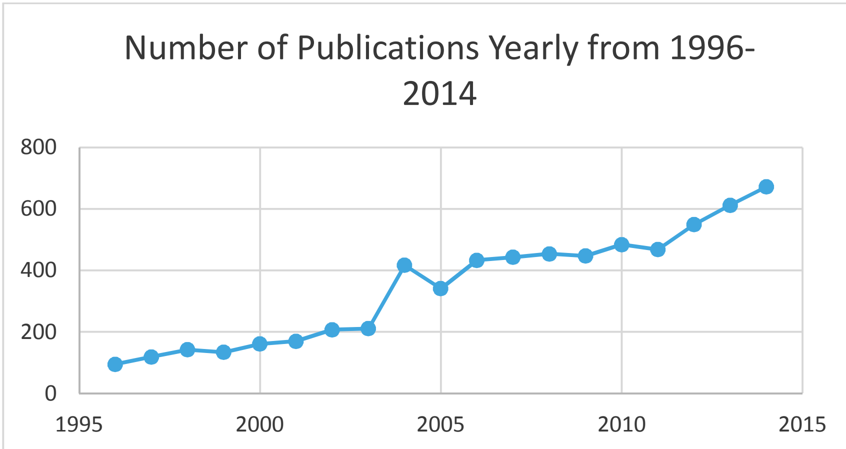
There are also examples of text mining used with scientific publications and visualization of results (Erten et al, 2004; Faisal et al, 2007; Fox et al, 2006; Synnestvedt et al, 2005). Stapley and Benoit constructed a prototype for genome information retrieval from 2,524 documents and visualization, linking terms of statistical significance (Stapley & Benoit, 2000). The authors state the graphical representation offered by their prototype provides researchers with the ability to intuitively assess the information presented and determine its value. A drawback they identified was that the volume of information represented by the combination of colors, graphics, and codes allowed the user to see the structure of the links but not the content. Allendoefer et.al. (Allendoerfer et al, 2005) use bibliometric analysis to create a visualization using nodes and clusters of networks showing similar nodes. They state that the layers created by their visualization have the potential to hide data, particularly if the user is not familiar with the database used. Andronis et al. (Andronis et al, 2011) describe several studies using ontologies with literature mining and visualization of results (primarily heat maps and graphs) for drug repurposing applications (new uses for existing drugs). They conclude that biomedical literature mining is an effective technique to generate hypotheses for drug repositioning, and clustering algorithms to bring similar concepts visually together is an effective method to provide additional insight into potential discoveries.

We could find no reports of text mining of complete articles as part of the visualization literature. Nunes et al. describe BeCAS, the Biomedical Concept Annotation System, developed for biomedical concept identification of PubMed abstracts that link to the reference databases (Nunes et al, 2013). In the systematic literature review we conducted in 2013, we found many concepts in the literature. We eliminated the vast majority of the articles from our final set based on information from the abstracts. With the final set, we read each article; 47% of the final set were eliminated (West et al, 2014). This raises the question of the value of the articles we initially rated based on the abstract: Would our results have been different had we had the means to more closely evaluate the complete text of each? Can text mining be used to easily identify the relevant terms in the literature, and would visualization of the results from text mining more accurately define the visualization literature?

METHODS

We completed a systematic review of the literature in May-June 2013 using PubMed and Web of Knowledge (West et al, 2014) as part of our research evaluating visualization of large health related data sets. The PubMed review was supplemented using citation searching and gray literature searching. Reference lists from highly relevant articles were also reviewed to find additional articles. We restricted our literature search to articles published since 1996. A query using PubMed for “information visualization” returned 6,559 articles, surprising since information visualization is a fairly recent approach in health care research and not part of the lexicon for PubMed. As Figure 1 shows, the number of articles on information visualization has grown significantly since 1996.

Figure 1. The number of publications in a PubMed search using the term “information visualization”. Adapted from Results by Year: http://www.ncbi.nlm.nih.gov/pubmed/?term=information+visualization.



In conducting the literature search, we wanted to find articles about the use of EHR data using innovative visualization techniques or describing techniques that could be applied to EHR data. We define EHR data as data in electronic clinical records that contain clinical information (e.g., demographics, problems, treatments, procedures, medications, labs, images, providers) collected over time that can be shared among all authorized care providers. We define innovative visualizations as visualizations other than standard graphs traditionally used for displaying health care information (e.g., bar charts, pie charts, or line graphs) that use complex data, which we define as data with multiple types of variables and many data points resulting in an exceptionally large amount of data, such as that in an entire EHR system. Interaction with the data is a key characteristic of information visualization, e.g. zooming or mouse-over to show features of the visualization. We were interested in articles describing any innovative visualization techniques for vast amounts of information that might be the foundation for an interactive system, however, so also included articles describing static visual representations of large amounts of EHR data. Articles were excluded if they related to animals or plants, were position papers describing the need for visualizing data or ideas for techniques in visualization, or did not describe specific techniques used for the visualization or have figures showing the results from visualization (West et al, 2014).

A total of 847 references were finally retrieved from our initial search of PubMed and Web of Knowledge. After a search of the gray literature and hand-searching references from articles, an additional 44 papers resulted in 891 articles to begin the review with. Using the abstracts from the articles, we found the literature replete with articles on visualization in genetics, syndromic surveillance, and geospatial environmentally-aware data. There were many articles on the technical details related to visualization techniques. We excluded 666 articles because the visualizations discussed were diagnostic, did not relate to EHR data, focused on animals or plants, used genomics data, discussed geospatial data or syndromic surveillance, were position papers suggesting the need for visualization or describing a potential visualization technique, or were primarily discussions of the technical details of visualization (West et al, 2014).

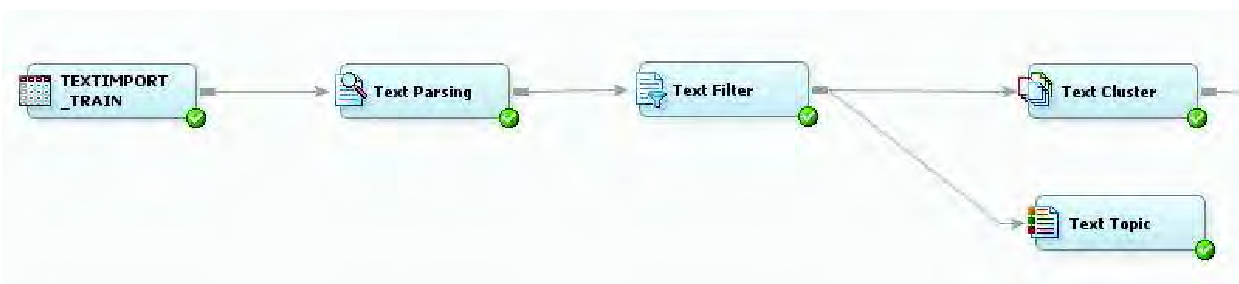
Sample

To assure we were familiar with the articles for this research using text mining, we drew from the literature we knew would be most similar to our interests in interactive health care information visualization. For our prototype and the visualization results we discuss later, we used the PDF files of 70 articles; this group of articles was included in the final abstract review for our systematic literature review in 2013. Our second experiment includes an additional 321 relevant articles, or a total of 391 articles. This combined group of 391 documents represents publications on visualization methodologies, tools, and applications to EHR data. The PDF of each document was used for this research. For our final exploration, we plan to use all 891 articles that were included in our systematic literature review conducted in 2013.

Text Mining

SAS Enterprise miner is the primary software used to process the documents. The text mining functionality of this software includes the ability to convert a collection of Adobe pdf documents into a SAS data set and then to parse, filter, and analyze the textual data, grouping similar documents as clusters and similar terms within the documents as topics. The experimental process sequence is shown in Figure 2.

Figure 2. Experimental Methodology Process Flow. Source: SAS.



The first node in Figure 2 is the conversion of the documents to a SAS data set. The SAS data set captures several properties (size, file location, etc.) of the document, and the document text is represented in a single variable by one long string.

The parsing node performs a number of processing functions on the text variable to identify a collection of terms in each document. Terms can be a single word or a multi-word expression. The single word terms can have parent-child relationships that account for synonyms and stemming. For example, stemming would establish a single term “zoom” to include the following: zooms, zooming, and zoomed. Synonyms can be identified from a database of predefined synonyms or by a user defined set. For example teach, instruct, educate, train could be a single document term. The basic concept is that parsing converts the string of text into a collection of terms that can be expressed as a term-by-document frequency matrix (see Table 1 for a simple generic example). The entries in this matrix are the raw counts of the number of occurrences of each term in each document. This data structure (with some refinement) is the basis for determining the relative degree of document similarity.

Table 1. Example of terms-by-document matrix: Adapted from Source: SAS.

TERM	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5	DOC 6	DOC 7	DOC 8	DOC 9	DOC 10
color	4	5	1	3	2	1	0	3	2	1
heat map	2	3	0	1	0	2	0	1	0	0
graphs	1	2	2	0	2	3	1	0	1	2
information	0	2	4	2	3	1	2	4	1	2
medical	1	2	0	1	3	2	1	0	2	1
chart	1	2	0	0	1	0	0	2	2	0
modeling	0	1	0	0	1	0	1	1	0	0
computer	1	2	0	1	3	0	0	1	2	2
interface	2	2	3	1	1	3	2	1	0	0
geographic	0	1	2	1	0	1	0	2	1	1
genes	1	0	2	1	3	2	1	0	3	1
phenotype	4	0	0	1	1	3	0	1	1	1

The matrix generated in this research is large and sparse with 20,000 terms (rows) and 391 documents (columns). There is an obvious need to reduce the number of terms and the dimensionality of the matrix prior to analysis. This is accomplished by the text filter. The filter node eliminates very common terms like “the”, the most frequently used word in the English language, and also eliminates terms rarely used in the document collection.

The text filter node enhances the information content of the raw term by a document frequency matrix, first weighting the raw frequency counts to reduce the effect of frequently occurring terms and then weighing the resulting terms to establish their ability to discriminate between documents in the collection being analyzed. The weighting function used to transform the raw frequency counts is a log function defined in Equation 1. The raw frequency counts, *fij* are the total number of occurrences of term i in document j.

$$g(f_{ij}) = \log_2(f_{ij} + 1) \quad (1)$$

The log weighting function dampens but does not eliminate the effect of terms that occur many times in a document. Alternatively, a binary function can be used to completely eliminate the effect of multiple term occurrences. The resulting matrix would then have an entry of 1 if there are one or more occurrences of the term and 0 if there are no occurrences.

The terms that are most effective at categorizing documents in a collection are those terms that occur in only a few documents but many times in those documents. The term weights are designed to identify these terms. There are two term weighting algorithms that are relevant for this research, entropy and inverse document frequency. Entropy is an information theory construct defined as follows where the entropy weight for term i, *wi*, is a function of the frequency, *fij,* the number of times the term i occurs in the document collection, *gi*, and *n*, the number of documents in the collection.

$$w_i = 1 + \sum_j \frac{\left(f_{ij/g_i}\right) * log_2(f_{ij})/g_i}{log_2(n)} \quad (2)$$

Inverse document frequency weights are determined as follows where *P(ti)* is the proportion of documents that have the term i.

$$w_i = log_2\left(\frac{1}{P(t_i)}\right) + 1 \quad (3)$$

A weighted term-by-document frequency matrix is calculated by first applying the frequency weighting function to the raw frequencies (Table 1) and then scaling that result by the term weighting function. Table 2 shows the transformation of the matrix in Table 1 by applying the log frequency weight function and the entropy term weight function.

Table 2. Term-Document Frequency from Text Miner. Term Weight=Entropy, Frequency Weight=Log. Adapted from Source: SAS.

Term	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5	DOC 6	DOC 7	DOC 8	DOC 9	DOC 10
color	0.271	0.301	0.116	0.116	0.116	0.116	0.116	0.116	0.116	0.116
heat map	0.114	03447	0.265	0.265	0.296	0.114	0.114	0.1144	0.114	0.114
graphs	0.046	03079	0.079	0.079	0.079	0.183	0.183	0.183	0.125	0.125
information	0.000	03502	0.000	0.202	0.202	0.320	0.320	0.202	0.523	0.470
medical	0.000	0.000	0.397	0.397	0.000	0.000	0.000	0.397	0.000	0.000
chart	0.522	0.000	0.522	0.000	0.000	0.000	0.000	0.000	0.000	0.522
modeling	0.000	0.522	0.000	1.522	0.000	0.000	0.000	0.000	0.522	0.000
computer	0.000	0.000	0.522	0.000	0.000	0.522	0.000	0.522	0.000	0.000
interface	0.000	0.000	0.000	0.000	0.522	0.000	0.000	0.528	0.528	0.000
geographic	0.723	0.000	0.000	0.000	0.000	0.000	1.146	0.000	0.000	0.000
genes	0.698	0.000	0.000	0.000	0.000	0.000	0.698	0.000	0.000	0.000
phenotype	0.000	0.000	0.000	0.698	0.698	0.000	0.000	0.000	0.000	0.000

Higher term weights are associated with terms that are more effective at categorizing documents. A cutoff value for the term weight is used to eliminate ineffective terms and reduce the dimensionality of the weighted term-by-document matrix. The analysis of the document collection begins with cluster analysis using hierarchical or expectation maximization algorithms. In preparation for cluster analysis, Singular Value Decomposition (SVD) is used to reduce the dimensionality of the input data (Trefethen et al, 1997). SVD is similar to principal components analysis and factors the weighted term-by-document frequency matrix into two orthogonal matrices U and V and a diagonal matrix ∑.

$$\mathbf{A} = \mathbf{U}\sum\mathbf{V} \tag{4}$$

SVD calculates only the first k columns of these three new matrices. Higher values of k explain more of the variance in the document collection. K must be large enough to capture the meaning of the document collection but not so large that it captures the noise. The projection of the original weighted term-by document matrix is projected onto the first k columns of U. Each row (term) of the term-by-document matrix is projected onto the first k columns of V.

The text topics and text clusters nodes characterize the text document collection. Each and every document is assigned to one and only one text cluster base on a distance metric calculated by Wards algorithm.

Experimental Design

Using our prototype with 70 documents, our second experiment using 391 documents, and third experiment using 897 documents, four document analyses per experiment explore the capabilities of machine learning methods to process and categorize documents. We will characterize the differences in results between the two term weighting algorithms: (1) entropy and (2) inverse document frequency. Each is applied to two different sets of document terms: (A) those selected by the machine algorithm from natural language and (B) a dictionary of terms defined by experts in health care visualization. Results from the four analyses (1A, 1B, 2A, and 2B) will each generate a complete set of results that include cluster assignments and document collection topics. These results will be analyzed using evaluation methods and a quantitative assessment.

An evaluation of the clusters and topics generated will be conducted using a panel of three experts in health care visualization, people who are the likely users of the technique we propose. A questionnaire to guide the panel of experts' feedback will be used to evaluate the panel's perceptions and opinions on the adequacy, effectiveness, and usability of results from each of the four cases. Results will be correlated and single metric measures like the Rand index or F-measures will be used to score the four experiments.

These results will be assessed quantitatively by a random subset removal technique. This consists of five trials where a random subset of the documents is removed from the data set. The analysis is then re-run for all four experiments and the percentage of documents in each cluster is contrasted in the before and after subset removal cases. If the clusters fundamentally represent the information in the document collection, the subset removal should not create major changes in cluster size or composition.

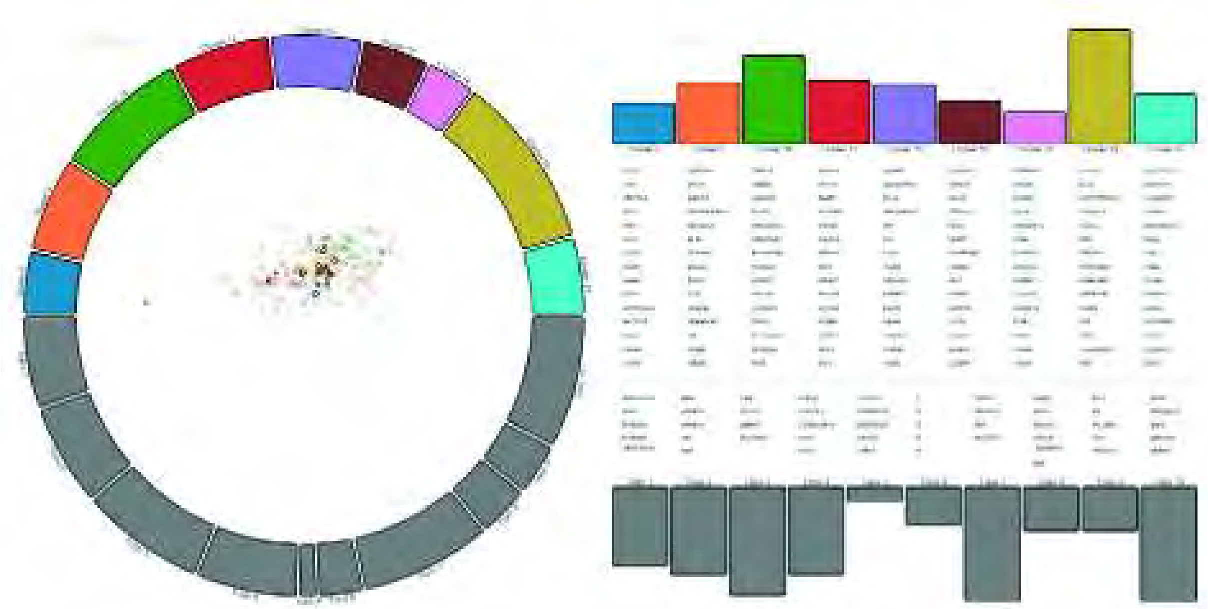
RESULTS AND DISCUSSION

This research is in progress, therefore our results section is not complete. The findings below include (1) the qualitative analysis of 70 documents and application of the findings for the prototype interactive visualization and (2) the quantitative analysis of the our second experiment using 391 documents .The expert panel evaluation of the and interactive visualization will be complete by the time of the DSI annual meeting.

Prototype and Interactive Visualization

For our prototype, we included 70 documents from the health care visualization literature as described earlier. Using results of the qualitative analysis as previously described, text was converted to numerical data using entropy. We have developed an interactive visualization tool using the D3 (Bostock et al., 2011) JavaScript library to display the resulting clusters, terms, and topics. Results of the visualization are shown in Figure 3. By visualizing these clusters and topics, we can evaluate those prevalent in the literature and those missing.

Figure 3. Document View (left) and Term View (right) linked visualizations of clusters and topics in the health care visualization literature.



In the Document View circle on the left, the top half are clusters represented by colored arcs; the size of each arc is proportional to the number of documents in the cluster. Gray arcs in the bottom half of the Document View circle are topics, with size proportional to number of documents with that topic. Inside the Document View circle is a scatter plot. Documents are displayed as pale-colored circles, with each color a cluster and the size of the circle proportional to the number of topics. Documents belong to a single cluster but may have multiple topics (or no topics). Colored rings are cluster centers with the radius proportional to the number of documents in the cluster. Black rings are topic centers with the radius proportional to the number of documents with that topic. In the Term View to the right of the Document View, the colored bars on the top are clusters, with height proportional to the number of documents in that cluster. The gray bars on the bottom are topics, with height proportional to the number of documents with that topic. Terms, taken from cluster descriptions and topic terms are noted in text beneath/above each cluster/topic. It is easy to determine that Topic 7 and 10 have the greatest number of documents, and Topic 5 has the least. Cluster 19 has the most documents and Cluster 18 has the fewest.

Interaction with the visualizations provides the user with a quick and easy way to gain a great deal of information about the literature. The user can explore the literature about the topics and articles published to date by mouse clicking various sections of the visualizations. It is possible to identify closely related clusters themes identified by the topic, their relationship to each other, the clusters and topics shared with common terms, the terms that comprise each cluster of documents, and the various authors of associated documents (Figures 4 and 5). This is a powerful way to determine like topics, the authors associated with a particular cluster who have published on specific topics, terms that might be used in a more targeted literature search on a given topic, and what topics have the greatest or least number of publications. The results from visualization of the output from machine learning applied to documents has the potential to become a useful way to determine what the literature contains, identify publications that might be similar to those the user is already familiar with, and identify like researchers in a given field and their area of concentration.

Figure 4. Mouse-over of Topic 10 (far right) displays closely related cluster and topic themes (terms), denoted by the box around each.

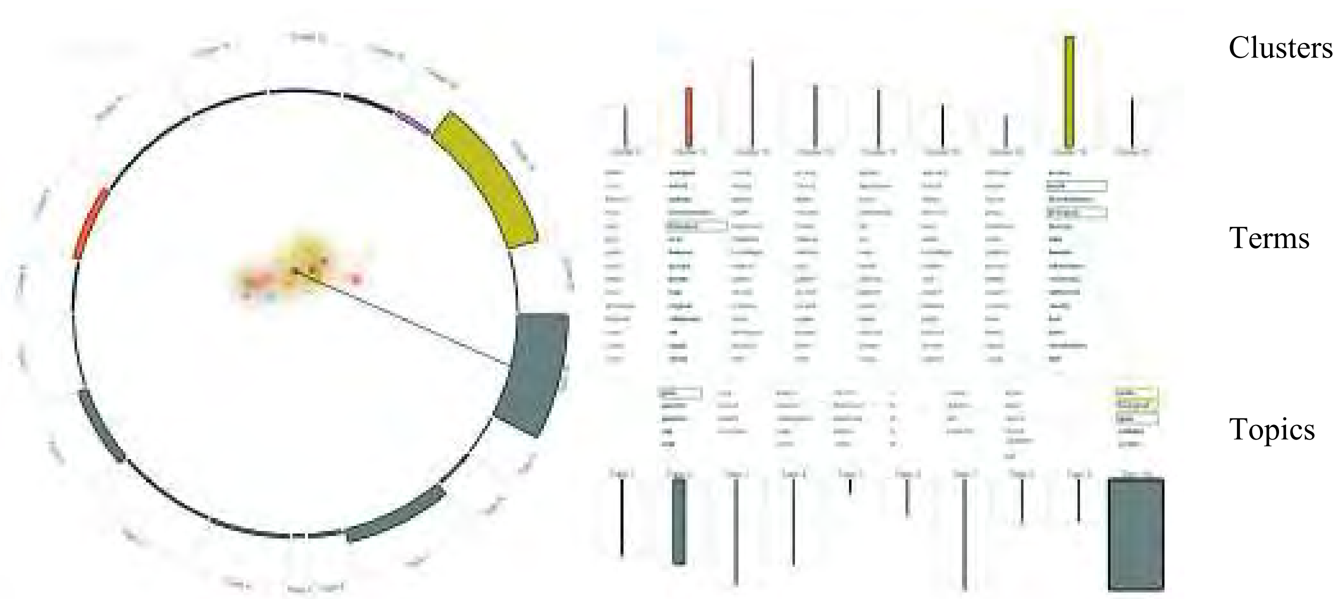
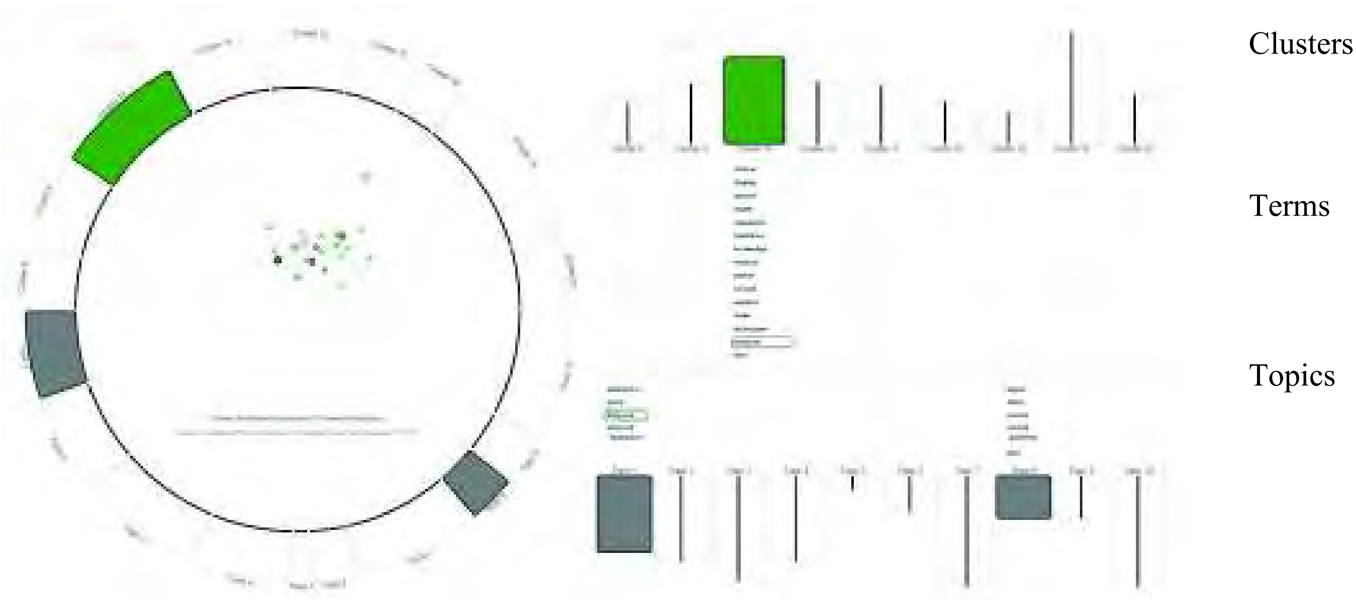


Figure 5. Two authors and their publication titles are displayed in the inside bottom of the Document View circle when highlighting a cluster that also displays a related topic and terms.



Document Collection Analysis

Our second experiment includes 391 documents. The output for one of our four analyses is described below. This analysis uses log weighting for terms, the inverse document frequency algorithm, and a set of 373 terms specific to the field of healthcare visualization developed by experts in the field.

Singular value decomposition transforms the document collection from text into a vector of numbers that can be analyzed by quantitative algorithms. Documents are clustered into 11 groups based on the distance between the documents. The results of Ward's hierarchical clustering algorithm are shown in Table 3. The (+) sign in front of some of some of the cluster description terms indicates the presence of synonyms or stem terms. The number of documents in each cluster is listed in the Frequency column and shown graphically in the bar chart of Figure 4.

Table 3. Clusters of similar documents.

Cluster ID	CLUSTER DESCRIPTION TERM	Frequency
9	fisheye spatial metadata 'computer graphics' +shape 'information visualization' visualization information visualizations +mapping multidimensional +focus +size interactive +cost	44
10	geographic spatial environmental +frequency space +shape +complexity +size times +space +evaluation novel computational +focus +overview	32
11	mappings environmental geographic 'user interfaces' system spatial multidimensional +flow +mapping visualizing phenotypes +phenotype costs times 'time-oriented clinical data'	7
13	genomics phenotypes +phenotype large-scale +visualization novel +information +system +scale +size technologies +complexity systems environmental times	34
19	'time-oriented data' temporal intelligent multidimensional 'information visualization' spatial visualization interactive visualizing 'data mining' 'data analysis' +aggregation 'time-oriented clinical data' +complexity information	15
20	genomics metadata computational +shape multidimensional large-scale +zoom 'data visualization' +scale +size +mapping 'data analysis' +visualization +space	34
22	'time series' +zoom +size 'data mining' +technology large-scale +frequency 'data analysis' +space +visualization +cost +information +scale visualizations +overview	17
24	genomics novel computational +mapping multidimensional +visualization programming +information large-scale +size 'data analysis' +space technologies +complexity 'data mining'	83
25	'knowledge base' intelligent 'time-oriented clinical data' temporal 'time-oriented data' +aggregation 'medical informatics' 'data mining' evaluation 'information visualization' 'time series' knowledge multidimensional computational spatial	28
26	temporal healthcare 'health care' 'medical informatics' +timeline 'decision making' +cost intelligent time +evaluation information +overview times +system 'information visualization'	75
28	evaluation 'data analysis' novel costs technologies +cost design +flow effectiveness +technology 'information visualization' environmental geographic visualization knowledge	22

Figure 6. The number of documents within Document clusters labeled with Cluster ID Number

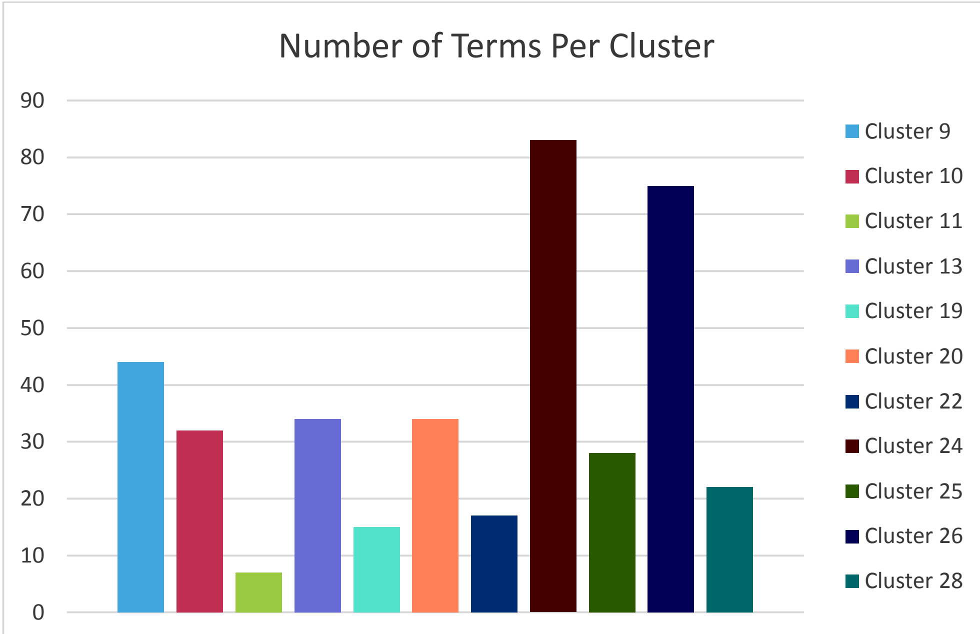
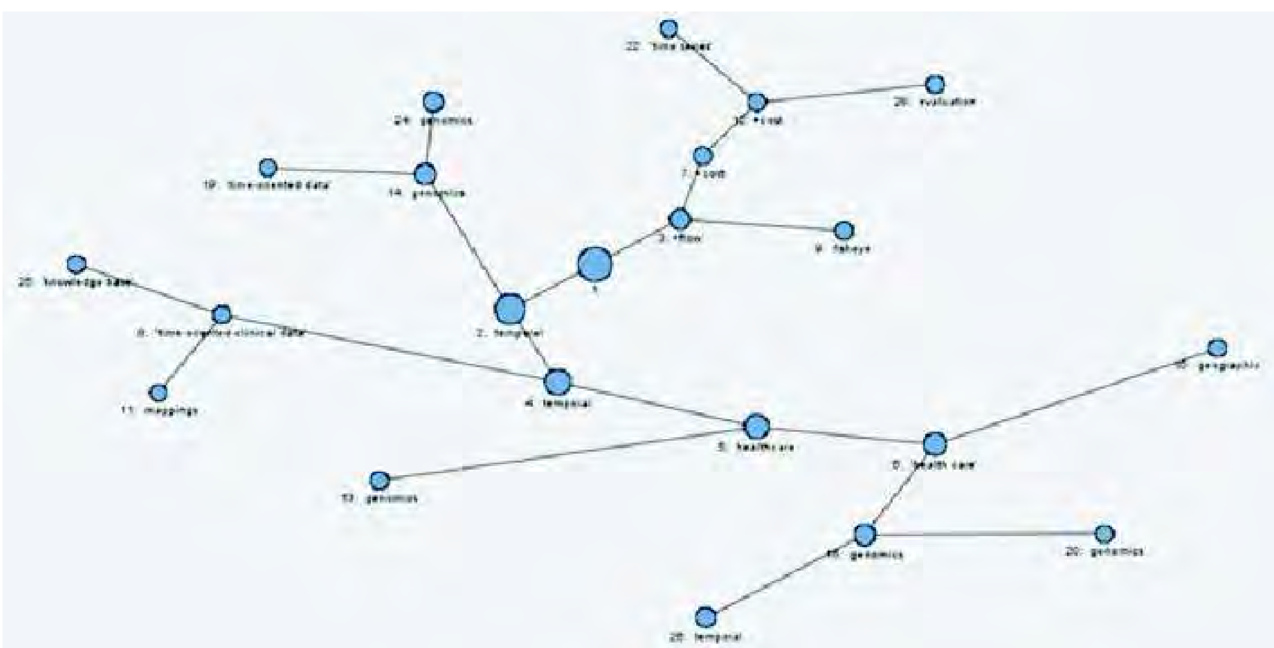


Figure 5 shows the hierarchical development of the document clusters. The largest circle, Cluster 1, is the entire collection of 391 documents. The collection is first split into two groups. The larger, identified as Cluster 2, is to the left and consists of 308 documents distinguished by the labels temporal visualization and genomics. The smaller cluster to the right of Cluster 1 is Cluster 3 and has 83 documents identified as flow and spatial visualization. Successive levels of cluster formation lead to the terminal clusters in Table 3.

Figure 7. Ward's cluster hierarchy.



The text mining algorithm also extracts topics from the document collection. Topics are common themes that occur as subsets of the document collection. The topics identified by the text mining algorithm for this experiment are summarized in Table 4.

Table 4. Document collection topics

Topic #	Topics				
1	"+visualization	+information	+system	+time	+technology"
2	"time-oriented data	intelligent	knowledge base	time-oriented clinical data	semantics"
3	"information visualization	design	+cognition	visualization and computer graphics	visualization"
4	"healthcare	healthcare	healthcare	health care	biomedical informatics"
5	"geographic	spatial	environmental	environmental	spatial"
6	"event sequences	visual analytics	EHR	visualization and computer graphics	+frequency"
7	"multidimensional	data visualization	+dimensionality	+mapping	+space"
8	"genomics	genomics	heatmaps	visual analysis	genes"
9	"metadata	biomedical informatics	metadata	+saturation	focus"
10	"+phenotype	+aggregation	genomics	intelligent	+dimensionality"
11	"time series	time series data	information visualization	dynamic query	time"
12	"fisheye	+zoom	+focus	zoom	fisheye"
13	"temporal	temporal	temporal	+timeline	artificial intelligence"
14	"environmental	+saturation	+flow	simultaneous	geographic"
15	"EHR	medical informatics	electronic health record	+timeline	EHR systems"
16	"imaging	medical imaging	spatial	grayscale	+flow"
17	"+precision	precision	+saturation	+vision	+frequency"
18	"+cost	health care	information technology	healthcare	+technology"
19	"exploratory data	data analysis	novel	programming	knowledge"
20	"medical informatics	decision making	decision support	EHR	artificial intelligence"

Results of document analyses will be presented to a panel of experts to assess the relative effectiveness of machine-based text mining analysis of research documents. Once there is concurrence on the effectiveness of the clusters and terms, we plan to conduct a third experiment using all 891 documents that comprised the initial set of documents from our 2013 literature review. These results will then be visually represented following the output of our prototype.

ACKNOWLEDGEMENTS
The literature review described in this paper was supported with funding from the US Army Medical Research and Material Command (USAMRMC), grant number W81XWH-13-1-006.

REFERENCES

Agarwal, P., & Searls, D.B. (2008). Literature mining in support of drug discovery. *Briefings in bioinformatics*, *9*(6), 479-492.

Allendoerfer, K., Aluker, S., Panjwani, G., Proctor, J., Sturtz, D., Vukovic, M., & Chen C. (2005, October). Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization. In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium, 195-202.

Andronis C., Sharma A., Virvilis V., Deftereos S., & Persidis A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, *12*(4), 357-368.

Bellis L., Akhtar R., AlLazikani B., Atkinson F., Bento A.P., Chambers J., & Overington J. (2011). Collation and data-mining of literature bioactivity data for drug discovery. *Biochemical Society Transactions*, *39*(5), 1365.

Bostock, M., Ogievetsky, V., & Hear, J. (2011). Data-driven documents*. IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2301-2309.

Bush G.W. (2004a) State of the Union Address, Promoting Innovation and Competitiveness, President Bush’s Technology Agenda.

Bush G.W. (2004b) Office of the Press Secretary, the White House. Executive Order: Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. Press release, April 27, 2004. http://www.whitehouse.gov/news/releases/2004/04/print/20040427-4.html2004 (accessed 23 Jul 2013).

Card S.K., Mackinley J.D., & Shneiderman B., eds. (1999) Readings in information visualization: using vision to think. Morgan Kauffman.

Charles D., Gabriel M.; & Searcy T., (2015). Adoption of Electronic Health Record Systems among U.S. NonFederal Acute Care Hospitals: 2008-2014. ONC Data Brief No. 23, April 2015. Accessed April 27, 2015 at http://healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf,

Chaussabel D. (2004). Biomedical Literature Mining. *American Journal of Pharmacogenomics*, *4*(6), 383-393.

De Bruijn B., & Martin J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. *International journal of medical informatics*, *67*(1), 7-18.

De Chassey B., Navratil V., Tafforeau L., Hiet M.S., Aublin-Gex A., Agaugue S., & Lotteau V. (2008). Hepatitis C virus infection protein network. *Molecular systems biology*, *4*(1).

Erten C., Harding P.J., Kobourov S.G., Wampler K., & Yee G. (2004). Exploring the computing literature using temporal graph visualization. *Electronic Imaging 2004, International Society for Optics and Photonics*, June, 45-56.

Faisal S., Cairns P., & Blandford A. (2007). Building for Users not for Experts: Designing a Visualization of the Literature Domain. *Information Visualization, 11th International IEEE Conference*, July, 707-712.

Fox E.A., Neves F.D., Yu X., Shen R., Kim S., & Fan W. (2006). Exploring the computing literature with visualization and stepping stones & pathways. *Communications of the ACM*, *49*(4), 52-58.

Hu X., Yoo I., Rumm P.,& Atwood M. (2005). Mining candidate viruses as potential bio-terrorism weapons from biomedical literature. In *Intelligence and Security Informatics*, Springer Berlin Heidelberg, 60-71.

Hur J., Schuyler A.D., & Feldman E.L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, *25*(6), 838-840.

Jung J.Y., DeLuca T.F., Nelson T.H., & Wall D.P. (2014). A literature search tool for intelligent extraction of disease-associated genes. *Journal of the American Medical Informatics Association*, *21*(3), 399-405.

Labaer J. (2003). Mining the literature and large datasets. *Nature Biotechnology*, *21*(9), 976-977.

Navarro,M.A., & Iratxeta, C.P. (2015). Text mining of biomedical literature. *Methods*, *74*, March, 1-106.

Nunes T., Campos D., Matos S., & Oliveira J.L. (2013). BeCAS: biomedical concept recognition services and visualization. *Bioinformatics, 29*(15), 1915-1916.

Papanikolaou N., Pavlopoulos G.A., Pafilis E., Theodosiou T., Schneider, R., Satagopam V.P., & Iliopoulos I. (2014). BioTextQuest+: a knowledge integration platform for literature mining and concept discovery. *Bioinformatics*,*30*(22), 3249-3256.

Poos K., Smida J., Nathrath M., Maugg D., Baumhoer D., Neumann A., & Korsching E. (2014). Structuring osteosarcoma knowledge: an osteosarcoma-gene association database based on literature mining and manual annotation. *Database, 2014*, 1-9.

Shatkay H., & Feldman R. (2003). Mining the biomedical literature in the genomic era: an overview. *Journal of computational biology*, *10*(6), 821-855.

Shetty K.D., & Dalal S.R. (2011). Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association,18*(5), 668-674.

Shneiderman B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *IEEE Symposium on Visual Languages*, September, 336-343.

Srinivasan P. (2004). Text mining: generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology, 55*(5), 396-413.

Stapley B.J.,& Benoit G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pacific Symposium of Biocomputing, 5*, January, 529-540.

Synnestvedt, M. B., Chen, C., & Holmes, J. H. (2005). CiteSpace II: visualization and knowledge discovery in bibliographic databases. *AMIA Annual Symposium, 2005*, 724.

Thieu T., Joshi S., Warren S., & Korkin, D. (2012). Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics*, *28*(6), 867-875.

Trefethen, L.N.; Bau III, D. (1997). *Numerical linear algebra.* Society for Industrial and Applied Mathematics, 361-369.

West, V. L., Borland, D., & Hammond, W. E. (2015). Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, *22*(2), 330-339.

Xiang, Z., Qin, T., Qin, Z. S., & He, Y. (2013). A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. *BMC systems biology*, *7*(Suppl 3), S9.

Appendix O

Survey on the Use of DEDUCE Queries

We are conducting an in-depth review of data queries to identify what data elements are included in queries, which will be used as a means to explore novel visualizations of large health data sets. We expect this approach to digitized healthcare data will lead to effective visualization of data, with an understanding of aggregated data that leads to discoveries within the data that would otherwise not be possible.

To help with this exploration, we hope you will be willing to share your thoughts and ideas with us about your use of the DEDUCE query system. Your responses will be aggregated with other respondents, and your name will not be used in any of the reports that might arise from this survey. Survey responses will be used to help us understand what information clinicians seek from data available to them, which will hopefully help us find the most effect ways to visualize the information.

1. Do you use DEDUCE for data queries?
   - Yes
   - No

2. Do you run your own queries?
   - Yes, always
   - Yes, most of the time
   - Sometimes
   - Seldom
   - No, never

3. Do you have someone else run the queries for you? If so, who?
   - No / Yes, followed by:
   - Clinical coordinator
   - Fellow
   - Clinical manager
   - Nurse Practitioner
   - Physician Assistant
   - Someone was trained to run all department's queries

4. Approximately how many times have you initiated a DEDUCE query in the past 2 years?
   - >20
   - 15-19
   - 10-14
   - 5-9
   - 1-4
   - 0

5. Why do you run queries? Check all that apply.
   - Thinking about writing a grant but need to know prevalence in Duke patient population
   - Need information for writing a grant

- Need to see if there are enough patients who will meet inclusion criteria to participate in an industry-sponsored clinical trial
- Searching for treatment methods
- Searching for outcomes
- For quality improvement
- Other clinical reasons: ____________
- Additional reasons other than clinical reasons:______________

6. What information do most of your queries seek? Select all that apply:
   - Demographics
   - Vital signs
   - Diagnoses
   - Medications
   - Procedures
   - Laboratory data
   - Imaging data
   - Device information
   - Encounters
   - Physicians
   - Other. Please list anything not on this list.

7. Do most of your queries provide you with the kinds of information you were looking for?
   - Yes, always
   - Yes, sometimes
   - No

8. Was the information you were seeking available with the first query that was run? If not, approximately how many times do you revise most of your queries before you get the information you want?
   - Almost always satisfied with the first query done
   - Every query usually needs to be revised
   - Usually revise once
   - Usually revise twice
   - Usually revise 3-4 times
   - Usually revise 5 or more times

9. In what format was the query information first presented to you?
   - Excel table
   - ASCII file
   - Bar graph
   - Line graph
   - XY graph
   - Other. Please list

10. Did you change the information to another format? If so, what did you use?
    - Excel table

- ASCII file
- Bar graph
- Line graph
- XY graph
- Other. Please list all you have used.

11. Would you be willing to provide feedback to us in the future regarding the usefulness of various ways to visualize data?
    - No
    - Yes
    - If yes, contact information:
        - Name
        - Email address
        - Phone number

Appendix P

APPENDIX A: Follow-up Questionnaire and Data from Evaluation of Various Data Visualizations

The chart below is raw data from the Questionnaire developed for use with an evaluation of various visualizations. It is followed by a graphical analysis of the three questions included with each graph: (1) Information is easy to find. (2) Information is easy to understand. (3) This is a useful way to look at the data.

Respondents	10		
Dates	June 2015		
Approximate Age	Mid 30s to early 60s		
Sex	Female (3); Male (7)		
BACKGROUND	MD(6); Pharmacist(1); Administrative Director(1); Data analyst(1); Project Leader(1)		
No. of Recorded Queries	4-177		
Self-reported number of queries last 2 yrs.	0-50; 5 over 12		
Use of DEDUCE queries in last 2 yrs	Quality Improvement (7); Outcomes (4); See if there are enough pts for study (2); Grant information (3); Utilization rates (1); Patient accession (1)		
Info pulled using DEDUCE	Encounters(6); Medications(6); Demographics(5); Physicians(5); Procedures(4); Lab Data(4); Diagnoses(4); Geographic data(3); Imaging Data(3); Social data(2); Vital signs(2); Other(1:only counts)		
Change info and how?	3: Bar graphs (3); Line graphs (2); Pie chart (1); Scatter plot (3)		
	Likert Scale: (1) best; (5) least favorite		
Length of Stay		Range	Avg
1.1.1.Bar graph	The information is easy to understand	1-2	1.8
	I can quickly find the information I am looking for	1-5	2.8
	This is a useful way to look at the data	2-4	2.6
1.1.2.Box& Whisker	The information is easy to understand	1-5	3
	I can quickly find the information I am looking for	1-5	3.2
	This is a useful way to look at the data	1-5	3.1
1.1.3.Bubble	The information is easy to understand.	2-5	3
	I can quickly find the information I am looking for	2-5	3.1
	This is a useful way to look at the data	1-5	2.9
1.1.4.Scatterplot Distribution	The information is easy to understand	1-4	2.4
	I can quickly find the information I am looking for	2-5	3.2
	This is a useful way to look at the data	1-5	3.6
1.1.5.Scatterplot	The information is easy to understand	2-5	3.8
	I can quickly find the information I am looking for	3-5	4.2
	This is a useful way to look at the data	2-5	4
RANKING	BAR GRAPH - Average number days of stay per DRG	1-4	1.7
	BOX & WHISKER PLOT- Average length of stay per DRG	1-5	2.8
	BUBBLE GRAPH - Total/DRG and Average Length Of Stay	1-4	2.8
	SCATTERPLOT DISTRIBUTION - Length of stay per DRG	2-5	3.2
	SCATTERPLOT - Length of stay per DRG over time	3-5	4.5
DRGs		Range	Avg

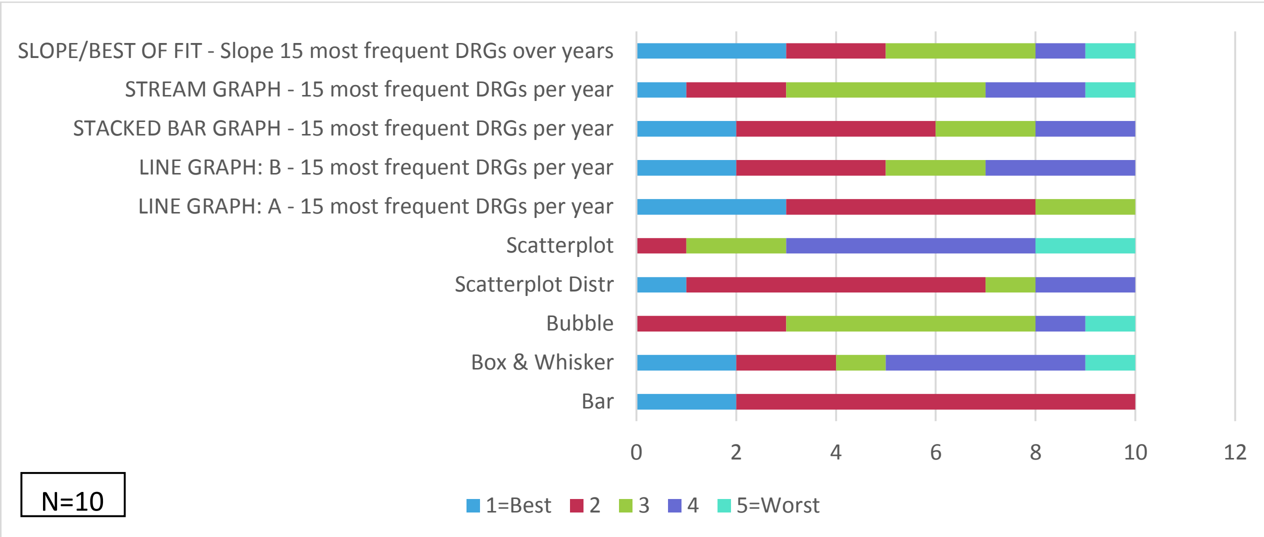
1.3.1.Line Graph A	The information is easy to understand	1-3	1.9
	I can quickly find the information I am looking for	2-4	2.5
	This is a useful way to look at the data	1-4	2.5
1.3.2.Line Graph B	The information is easy to understand	1-4	2.6
	I can quickly find the information I am looking for	2-4	2.8
	This is a useful way to look at the data	1-5	3.1
1.3.3.Stacked Bar Graph	The information is easy to understand	1-4	2.4
	I can quickly find the information I am looking for	1-5	2.8
	This is a useful way to look at the data	1-5	2.8
1.3.4.Stream Graph	The information is easy to understand	1-5	3
	I can quickly find the information I am looking for	2-4	3.3
	This is a useful way to look at the data	1-4	3.1
1.3.5.Slope/Best of Fit	The information is easy to understand	1-5	2.5
	I can quickly find the information I am looking for	1-5	2.6
	This is a useful way to look at the data	1-5	3
RANKING	LINE GRAPH: A - 15 most frequent DRGs per year	1-4	2.4
	LINE GRAPH: B - 15 most frequent DRGs per year	5-4	3.1
	STACKED BAR GRAPH - 15 most frequent DRGs per year	1-5	2.9
	STREAM GRAPH - 15 most frequent DRGs per year	1-5	3.6
	SLOPE/BEST OF FIT - Slope 15 most frequent DRGs over yrs	1-5	3
Change in VS		Range	Avg
2.2.2.Heatmap Carryover (Wt/pt/yr)	The information is easy to understand	1-5	2.9
	I can quickly find the information I am looking for	1-5	2.9
	This is a useful way to look at the data	1-5	2.8
2.2.3.Heatmap Carryover (Diastolic BP)	The information is easy to understand	2-3	2.6
	I can quickly find the information I am looking for	2-5	3.0
	This is a useful way to look at the data	2-5	3.1
2.2.4.Line Graph (Diastolic BP)	The information is easy to understand	2-5	3.9
	I can quickly find the information I am looking for	2-5	3.9
	This is a useful way to look at the data	1-5	3.9
2.2.6.Sankey Diagram	The information is easy to understand	2-5	4.1
	I can quickly find the information I am looking for	2-5	3.8
	This is a useful way to look at the data	2-5	3.7
RANKING	HEATMAP CARRYOVER: A - Average weight per patient per year	1-3	1.8
	HEATMAP CARRYOVER: B - Average diastolic blood pressure per pt per yr	1-4	2.3
	LINE GRAPH - Average diastolic blood pressure per patient per year	1-4	3
	SANKEY DIAGRAM - Relationships of increasing or decreasing vital signs	1-4	3.0
No. variables (2)		Range	Avg
3.2.1. Bar Graph	The information is easy to understand	1-5	2.9
	I can quickly find the information I am looking for	1-5	3.3

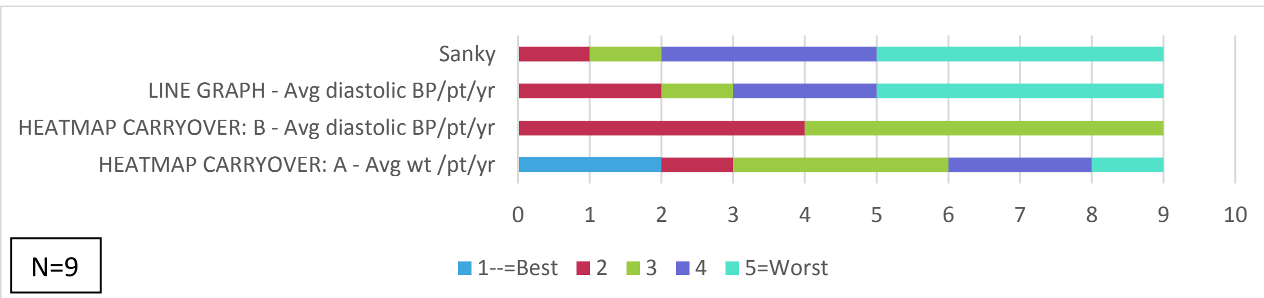
	This is a useful way to look at the data	1-5	3.0
32.2. Stacked Bar Graph	The information is easy to understand	1-4	2.3
	I can quickly find the information I am looking for	1-4	2.3
	This is a useful way to look at the data	1-4	2.3
3.2.3 Line Graph	The information is easy to understand	2-4	2.6
	I can quickly find the information I am looking for	2-4	2.7
	This is a useful way to look at the data	2-4	3.0
3.2.4. Bubble Graph	The information is easy to understand	2-5	3.4
	I can quickly find the information I am looking for	2-5	3.6
	This is a useful way to look at the data	2-5	3.7
3.2.5 Bipartite Graph	The information is easy to understand	2-5	2.9
	I can quickly find the information I am looking for	3-5	3.4
	This is a useful way to look at the data	1-5	2.9
3.2.6.Sankey Diagram	The information is easy to understand	1-4	2.8
	I can quickly find the information I am looking for	2-4	3.2
	This is a useful way to look at the data	1-4	2.7
3.2.7.Marimekko Chart	The information is easy to understand	1-2	1.7
	I can quickly find the information I am looking for	1-3	2.2
	This is a useful way to look at the data	1-4	2.17
RANKING	BAR GRAPH - Race & Diagnosis	1-7	3.9
	STACKED BAR GRAPH - Race & Diagnosis	1-7	3.2
	LINE GRAPH - Race & Diagnosis	2-7	4.3
	BUBBLE GRAPH - Race & Diagnosis	1-7	4.7
	BIPARTITE GRAPH - Race & Diagnosis	1-7	4.1
	SANKEY DIAGRAM - Race & Diagnosis	2-5	4.0
	MARIMEKKO CHART - Race & Diagnosis	1-5	3.2
No. variables (3)		Range	Avg
3.3.1.Nested Bubble Graph	The information is easy to understand	2-5	3.8
	I can quickly find the information I am looking for	2-5	4
	This is a useful way to look at the data	1-5	3.6
3.3.2.Sankey Diagram	The information is easy to understand	3-4	3.4
	I can quickly find the information I am looking for	3-4	3.4
	This is a useful way to look at the data	3-5	3.8
RANKING	NESTED BUBBLE GRAPH - Diagnosis, Gender, Race	1-2	1.6
	SANKEY DIAGRAM - Gender, Diagnosis, Race	1-2	1.6
	COMMENTS		
	1=Agree with statement 0=Do not agree with statement	% agree	
4: Interactive			
Bipartite Graph	I have seen this type of visual display before.	40%	
Sankey Diagram		40%	
Parallel Sets		20%	
Radial Coordinates		30%	
Bipartite Graph	It was easy to understand the data presented	60%	
Sankey Diagram		40%	
Parallel Sets		0%	
Radial Coordinates		0%	

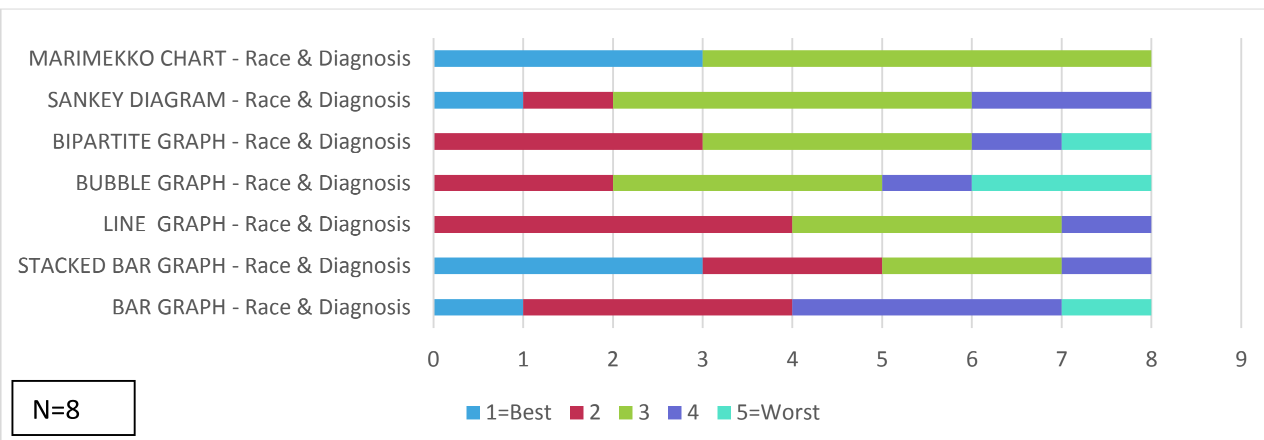
Bipartite Graph	This was not easy to understand at first but I did after looking at/talking about it.	40%
Sankey Diagram		50%
Parallel Sets		30%
Radial Coordinates		10%
Bipartite Graph	I do not understand this visualization.	0%
Sankey Diagram		20%
Parallel Sets		50%
Radial Coordinates		50%
Bipartite Graph	I can think of ways this visualization would be helpful.	80%
Sankey Diagram		60%
Parallel Sets		20%
Radial Coordinates		30%
Bipartite Graph	I would never consider using this type of visualization in my work but might recommend it to others.	0%
Sankey Diagram		30%
Parallel Sets		40%
Radial Coordinates		20%
Bipartite Graph	I would never recommend this type of visualization.	10%
Sankey Diagram		10%
Parallel Sets		50%
Radial Coordinates		30%

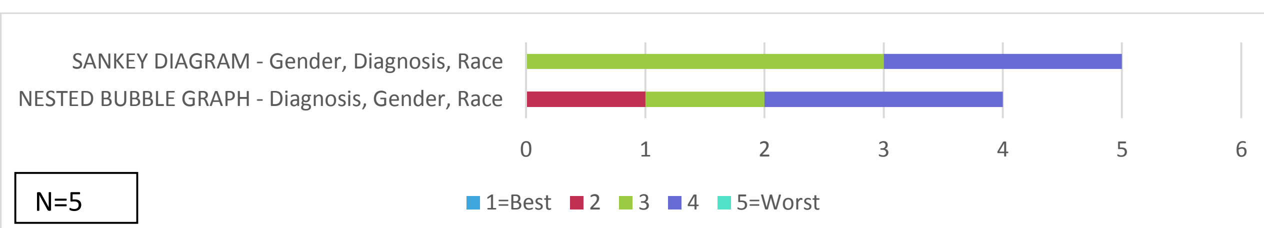
Information is easy to find.

Number of Responses: Likert Scale 1-5



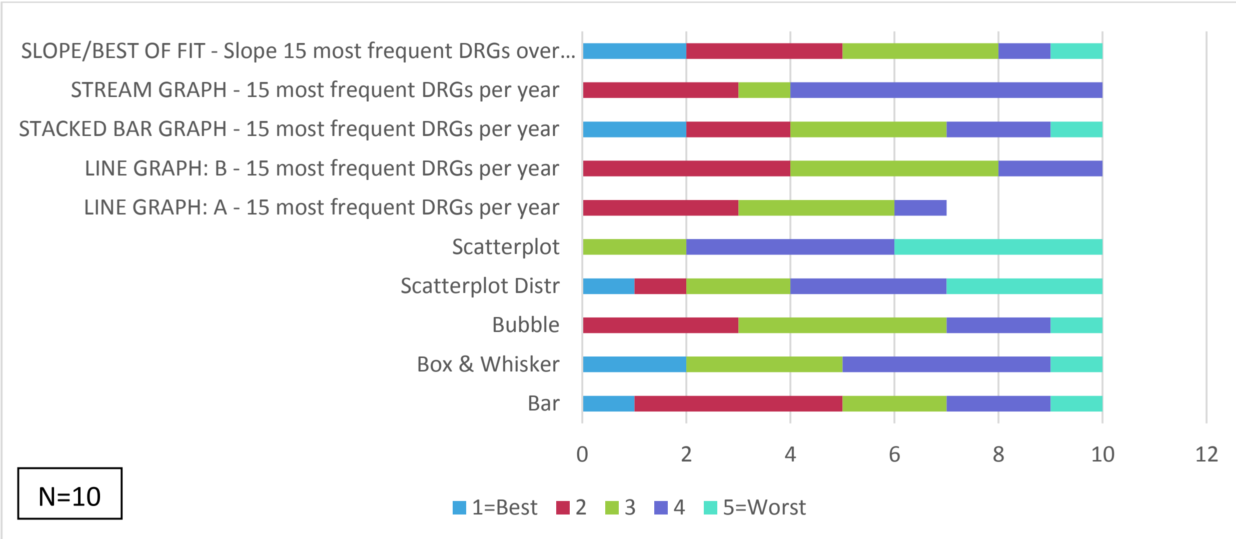


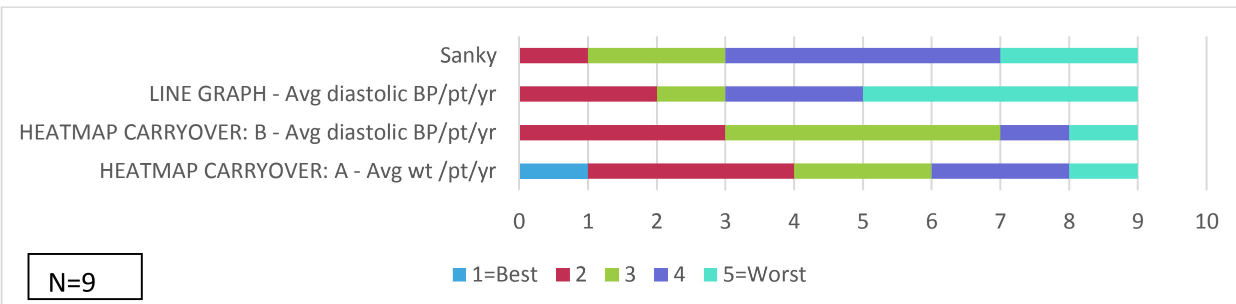


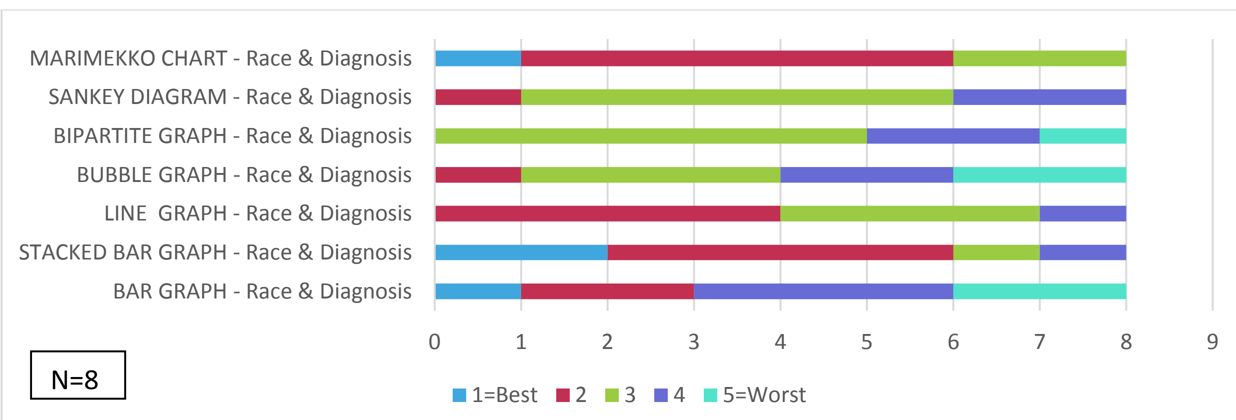


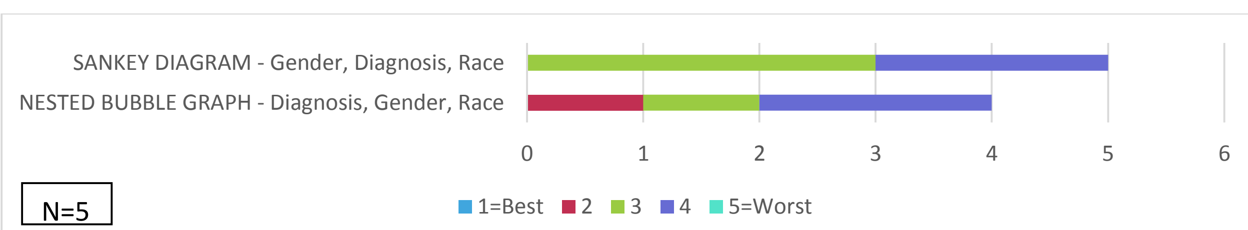
Information is easy to understand

Number of Responses: Likert Scale 1-5









This is a useful way to look at the data.

Number of Responses: Likert Scale 1-5